From Silicon Sampling to Population Steering: When Does Survey-Grounded Persona Generation Beat Demographic Prompting?

Jason Grey
jason@jason-grey.com

August 2026

Abstract

Products that simulate survey populations with LLM personas make three different promises: predict a person, reproduce a population, and steer that population. We tested all three against a real election survey (ANES 2020) and got three different answers. **Predicting a person** from invented traits cannot work: traits sampled from someone's demographics carry no information about them beyond the demographics themselves, and across seven LLMs and two decoding configurations, grounding never decisively beats a well-calibrated demographic prompt; its one clear win is rescuing the single badly miscalibrated model in the roster. **Reproducing a population** works but is not special: grounded personas cut distributional error to a third of naive repeated prompting, but asking the model for the distribution outright, prompting it to verbalize a spread of answers, appending one calibrated label to the plain prompt, or looking the answer up in the very survey data that trained our generator all do about as well. What the persona machinery genuinely adds there is honest variance: naive prompting collapses a population to one modal answer with noise. **Steering** works exactly as far as two measurable channels allow—the label the persona wears, which the LLM obeys, and the generator's learned coupling between attributes, which is weak in our current model—and both channels can be audited in advance without a single LLM call. We argue vendors of persona systems should publish those audits. Along the way, scoring each layer directly instead of trusting end-to-end results caught two silent decode bugs that had corrupted every generated persona, a run that failed 95% of its calls while reporting success, and an overclaim in our own draft. The practical conclusion: for reading distributions off well-surveyed demographics, prompting or a lookup table is the better tool; persona simulation must justify itself on finer-grained outcomes, populations the model does not already know, and above all on interaction between simulated people—which this paper deliberately does not test. Those three boundaries are also the forward program. For a *documented* individual the information argument inverts: measured attributes are the ones that can carry signal. For a population no survey has measured, prompting and lookups have nothing to return, while a generative joint model can still answer and stays auditable against the surveys that built it. And honest per-persona variance, useless for reading off a share, is the raw material of a believable simulated room. Sequel studies test each claim, held to the audits introduced here.

1 Introduction

Argyle et al. (2023) demonstrated that a large language model conditioned on socio-demographic backstories drawn from real survey respondents emits response distributions that track those of

the corresponding human subpopulations—a property they call *algorithmic fidelity*. The result seeded a fast-growing practice of "silicon sampling": using LLM personas as inexpensive stand-ins for survey respondents, focus groups, and experimental subjects (Park et al., 2024; Kang et al., 2025; Cui et al., 2025). It also seeded a design question that the original work leaves open. A demographic backstory is a one-line sketch; a person is not. If a persona were instead *grounded*—its personality, moral foundations, values, ideology, and religiosity sampled jointly from the empirical distribution of real respondents conditioned on those same demographics—would the silicon sample get closer to the humans?

This paper reports a two-study arc that answers the question twice, at two different levels of analysis, with opposite verdicts.

Study 1 (individual level): grounding fails, and must fail. We compare, for each of the same held-out ANES respondents, a plain demographic backstory against a grounded persona built from the same demographics, across seven base LLMs. Grounding loses in the production configuration, decisively on the well-calibrated models; at its strongest configuration the gaps mostly contract into the precision bound, and grounding still never decisively beats a well-calibrated baseline. The better vanilla baselines match a supervised model trained directly on the ANES demographic-to-vote mapping—with no ANES training at all. The reason is structural, not fixable by richer personas: traits sampled from someone's demographics carry no information about that person beyond the demographics (Proposition 1), yet the elicitor substantially obeys whatever label the persona wears, so a noisy invented label displaces the model's own good priors.

Study 2 (population level): grounding wins its preregistered comparison, and four baselines then shrink the win. The product claim is not individual prediction; it is that K sampled personas form a population with calibrated, honestly-dispersed, steerable response distributions. Testing that first forced an audit of the generative layer's own decoded outputs, which exposed two silent defects that had been corrupting every persona's political fields; repairing them and calibrating the decoder (cutpoints, dispersion matching, correlated noise) cut the layer's distributional error from 0.417 to 0.130 without retraining. So repaired, grounded personas beat repeated vanilla prompting decisively—a third of its error, with survey-design-aware intervals separated—and keep realistic within-cell variance where vanilla prompting collapses to a quarter of it (the failure documented by Bisbee et al., 2024). Then the baselines close in: asking the model for the distribution outright is better still; verbalized sampling (Zhang et al., 2025) ties the grounded arm at a fifth of the calls; a weighted lookup on the generator's own training data matches it; and a single calibrated ideology label appended to the vanilla prompt recovers most of the advantage. Steering, finally, is bounded by two channels we can audit in advance without an LLM: the label the persona wears (obeyed, and caricatured by the vanilla arm) and the generator's learned cross-field coupling (weak in the current artifact).

The through-line is a working rule: *score each layer directly, at the level of the claim it makes.* Applied to the generative layer it found the decode defects; applied to run logs it caught an experiment that silently failed 95% of its calls while reporting success; applied to our own draft it retracted an overclaim about persona coherence. Every audit in this paper is cheap, most are free, and all are specified fully enough to reimplement.

Our contributions:

1. A three-level evaluation of one persona system against a held-out probability survey, with the verdict flipping by level, and a proposition that makes the individual-level limit exact rather than empirical (§4).

2. Two free audits that persona systems can ship as disclosures: direct scoring of the generative

layer's decoded fields (which exposed two silent defects invisible in rendered text and end-to-end metrics), and a per-knob steering audit that decomposes steering into a rendered-label channel and a learned-coupling channel (§5.2, §5.6).

3. Post-hoc decoder calibration—quantile-matched cutpoints, dispersion matching, and correlated noise—that repairs population fidelity and per-persona coherence without retraining, with the joint-dependence cost of each step measured (§5.3).

4. A population-level comparison against four baselines that narrow the claim rather than flatter it: repeated prompting, direct distribution readout, verbalized sampling, and same-data statistical models, with design-aware uncertainty (§5.4, §6).

2 Related work

Our starting point is Argyle et al. (2023), who showed demographic backstories give LLM samples demographically-correlated response distributions, and the critical literature that followed: default liberal-educated response profiles (Santurkar et al., 2023), variance collapse in synthetic samples (Bisbee et al., 2024), caricature under persona conditioning (Cheng et al., 2023), and weak persona effects at the individual level (Hu and Collier, 2024; Taday Morocho et al., 2026). Individual-level gains have instead come from grounding in *measured* individual data—interviews (Park et al., 2024; Kang et al., 2025) or behavioral traces (Li and Conrad, 2026)—which our Study 1 complements from the other side: population-sampled depth cannot do what individually-measured depth does. On the population level, Cao et al. (2025) fine-tune LLMs toward survey response distributions where we calibrate the generative layer instead; Zhao et al. (2025) benchmark synthetic respondents as diagnostic instruments; Lutz et al. (2025) study prompt formats, whose effects we bound as small next to content; and Zhang et al. (2025) supply the verbalized-sampling pattern we adopt as a baseline. Miranda and Balbi (2025) argue LLM survey predictions should be benchmarked against supervised models, a standard we adopt and extend to same-data statistical baselines. Appendix C details the per-work relationships.

3 The study system: a population-grounded persona sampler

AnthroSim is a persona-generation system built around `fusion_model_v1`, a latent factor model fit over multiple U.S. survey and administrative datasets. Its contract is:

$$\text{ACS backbone} \;\rightarrow\; \text{posterior latent state} \;\rightarrow\; \text{decoded attributes.}$$

The population backbone is a 270,479-cell lattice built from American Community Survey PUMS, carrying calibrated joint demographics (age, sex, race/ethnicity, education, income, state, and derived profession fields). Over this backbone, an eight-dimensional latent factor model is trained on 4.87M sparse observations from mounted survey sources—the Cooperative Election Study (CES), the General Social Survey (GSS), CPS Voting and Registration, Big Five inventories (OpenPsych IPIP-FFM, AutoMoto), the Moral Foundations Questionnaire (MFQ-2), Pol.is deliberation aggregates, Voteview ideal points, and EPA EJScreen environmental indicators—with demographic-overlap anchors tying non-ACS rows into the shared latent space. Sampling proceeds by compiling user-specified criteria ("evidence") into a reweighting of the backbone and a conditioning of the latent posterior; a persona is one draw: an ACS cell, a latent vector (optionally perturbed by a *latent temperature*), and decoded attribute fields (party identification, ideology,

Big Five, moral foundations, religious attendance, generalized trust). The ANES is *not* among the model's training sources, making it a genuinely held-out criterion throughout.

For LLM-facing use, a sampled persona is rendered to a first-person profile text; the product path (`ProfileData.to_first_person_intro`) renders an interview-style introduction (Appendix A). Elicitation is a single-shot, in-character survey call shared by all experimental arms (Appendix A); arms differ only in the persona text supplied. Appendix F gives the full model specification—observation model, weighting, optimization, identifiability, missingness handling, and generation pseudo-code.

4 Study 1: individual-level fidelity

4.1 Preregistered design

Hypothesis (H1, locked before any run). Grounded personas produce higher algorithmic fidelity to real ANES 2020 responses—2020 presidential vote choice and 7-point party identification—than demographic-backstory prompting, holding the base LLM and conditioning demographics constant.

The comparison is within-respondent and same-model. Every test-set respondent is predicted twice: once by the **Baseline arm**—a first-person demographic backstory in the style of Argyle et al. (2023)—and once by the **Grounded arm**, in which the same demographics are supplied as hard evidence to the fusion sampler and the resulting persona (with its sampled ideology, Big Five, moral foundations, attendance, and trust) is rendered to text. Both arms then pass through an identical single-shot elicitor at fixed temperature. A free diagnostic "tabular" readout logs the fusion model's directly-decoded `party_id` and `ideology` fields with no LLM call.

Leakage rule. Both arms must predict the held-out targets from non-target inputs. Conditioning demographics exclude vote and party ID; the grounded persona may surface inferred ideology and psychographics (that is the enrichment under test), but the literal scored targets never appear in any persona text. Vote and party are held out as a single political block. The grounded ideology is the fusion model's *inference from demographics alone*, not the respondent's self-report, so it does not constitute leakage—a distinction that matters for the circularity critique in the persona-prompting literature.

Ground truth. The ANES 2020 Time Series Study: 8,280 respondents, of whom 6,222 have a valid post-election weight, a valid 7-point party ID, and a major-party-or-other presidential vote. All variable codings were verified against the codebook (age `V201507x`, sex `V201600`, race/ethnicity `V201549x`, education `V201511x`, income `V201617x`, state `V203000`, party ID `V201231x`, vote `V202110x`, weight `V200010b`). Sample marginals match known ANES 2020 patterns (self-reported-voter vote share: 56% Democrat / 41% Republican / 3% other; party ID bimodal). All scoring is survey-weighted.

Metrics. (i) Individual fidelity: vote accuracy (later two-party *balanced* accuracy, for class-skew robustness), 7-point party-ID exact accuracy, MAE, and individual correlation. (ii) Subgroup-distribution fidelity: correlation and MAE of predicted vs. real cell-level shares across demographic cells. (iii) Calibration guardrails: absolute Republican-share bias and mean party-ID bias. The preregistered decision rule required the grounded arm to win both individual accuracy and subgroup correlation, for both targets.

4.2 Pilot and rendering controls

The pilot ran N=200 stratified ANES voters through both arms on `gpt-4o-mini`. The arms split along a discrimination–calibration axis, and H1 failed its decision rule: the vanilla baseline won every discrimination metric (vote accuracy 0.680 vs. 0.571; party-ID individual correlation 0.323 vs. 0.144; subgroup rank correlation 0.72/0.81 vs. 0.36/0.11), while the grounded arm won every calibration metric, most sharply on aggregate party-ID bias (0.03 vs. 0.97). The baseline's error is the liberal-educated default profile of Santurkar et al. (2023), measured: against a true 113 Democratic / 77 Republican split it predicted 158/42, a full point too Democratic on the 7-point scale. Grounding corrected that aggregate bias while adding individual noise—baseline points sit below the diagonal but well-ordered, grounded points centre on the diagonal but scatter (Appendix D, with the full pilot table).

Before trusting that, we asked whether the grounded arm was merely worded badly. Sixteen renderer variants (Appendix E) were swept with the persona fixed and only the text varying, with paired bootstrap intervals and a held-out validation split. Three results survived: *wording moves individual fidelity only within noise* (the best variant, an interview transcript consistent with Lutz et al., 2025, gained +0.014 balanced accuracy with a bootstrap bound touching zero, shrinking to $\Delta \approx +0.007$ under later cross-model checks); *content dominates*, since deleting the single grounded ideology line returns aggregate bias to the vanilla skew while every psychographic field combined moves almost nothing; and *temperature-0 completions drift between identical runs* by margins comparable to the effects being measured, so all subsequent experiments pool repeats. The second finding is the one Study 2 must eventually answer for: if one field carries the political behavior, the grounding may be a label injector rather than a joint model (§6).

4.3 Cross-model study on the repaired artifact

The pilot's calibration win was measured on one base model, and on a generative artifact later found to be defective (§5.2). We therefore report the cross-model comparison on the *repaired and calibrated* artifact in its production configuration, over seven base LLMs (N=150; N=60 for the most expensive), against a supervised reference fit on the same seven demographics (logistic regression and random forest, 5-fold CV over all 6,222 voters, scored through the identical pipeline). Table 1 and Figure 1 give the result; the pre-repair version of this table, which we no longer rely on, is in Appendix I.

Two facts reorder the conclusions. First, *good vanilla baselines match or exceed the supervised reference with no ANES training*: three models clear the 0.62–0.63 band. For demographics-only individual prediction, an off-the-shelf LLM with a plain backstory is at the practical state of the art (cf. Miranda and Balbi, 2025). Second, *grounding does not help, and where the comparison is decisive it hurts*: the baseline leads on five of seven models, by margins beyond the run's ±0.08 precision bound on four; the two small reversals and the llama gap are inside the bound, and we decline to interpret them—including the tempting story that the reversals are the worst-calibrated baselines being rescued, which fails its own check (one reversal's grounded arm has *worse* bias than its baseline, 0.82 vs. 0.52). A second rerun gives the grounded arm its *strongest* configuration—mode decode, where the repaired label carries its full 0.251 signal instead of the production configuration's honest per-draw noise—so that the negative result cannot be attributed to testing a deliberately noisy arm (Appendix D, Table 7). The baseline still leads five of seven, but most gaps contract into the precision bound; what survives interpretation is one decisive loss (claude-sonnet, −0.114) and one decisive win: gpt-4o-mini (+0.111), the most miscalibrated baseline in the roster (bias 1.32), genuinely rescued by a calibrated label ($1.32 \rightarrow 0.34$). Across

Table 1: Cross-model comparison on the *repaired* generative artifact (temperature 0; N=150 per model, N=60 for `claude-opus-4.8`*). Balanced accuracy / party-ID correlation / |mean party-ID bias|, survey-weighted; rows sorted by baseline accuracy. The backstory baseline leads on five of seven models, decisively on four; the two reversals (and the llama gap) are inside the ±0.08 precision bound of Appendix J and are not interpreted. The supervised reference on the same demographics is 0.620 (logistic regression) / 0.634 (random forest).

	Backstory baseline			Grounded persona		
Base model	bal. acc.	corr.	|bias|	bal. acc.	corr.	|bias|
claude-opus-4.8*	**0.764**	0.241	0.32	0.603	0.155	0.67
deepseek-chat-v3.1	**0.690**	0.329	0.23	0.475	0.066	0.81
claude-sonnet-4.6	**0.685**	0.254	**0.17**	0.583	0.129	0.70
llama-3.3-70b	**0.607**	0.231	0.35	0.570	0.155	0.61
gemini-2.5-flash	**0.597**	0.246	0.35	0.494	0.072	0.94
claude-haiku-4.5	0.573	0.206	0.52	**0.619**	0.093	0.82
gpt-4o-mini	0.523	0.220	1.36	**0.546**	0.040	**0.38**
Supervised reference (logreg / RF)	0.620 / 0.634			—		

both configurations the summary is: grounding never decisively beats a well-calibrated baseline, and its one decisive win is repairing the worst one.

A component ablation on the two extreme models (Appendix E) locates the effect inside the persona: psychographics without ideology is the *worst* configuration on both models, while on the well-calibrated model bare demographics beat every grounded variant. Grounded ideology is a crutch that rescues a broken baseline and dead weight on a good one—the same asymmetry Table 1 shows across models.

4.4 Mechanism: an information bound and an obedient reader

The failure has a compact information-theoretic explanation. Let X be the conditioning demographics, Y the respondent's held-out answer, and $Z \sim q(\cdot \mid X)$ the sampled persona attributes, drawn with randomness independent of the respondent given X.

Proposition 1 (No added information)**.** *If $Z \perp Y \mid X$, then $p(Y \mid X, Z) = p(Y \mid X)$. Population-sampled persona attributes carry no information about the respondent beyond the demographics they were sampled from.*

The proof is immediate from the definition of conditional independence; the content is in what it does *not* say. It bounds the *information* available, not the *performance* of any particular predictor. A finite, misspecified predictor conditioned on (X, Z) may do better or worse than the same predictor on X alone: better if Z happens to correct a bad prior, worse if it displaces a good one. Both outcomes appear in our data, and the mechanism decides which. Three measurements from the pilot's own logs—no new runs—assemble it:

1. **The grounded ideology label is individually near-random.** In the pilot (pre-repair artifact, mode decode), weighted correlation between the persona's sampled ideology and true 7-point partisanship was 0.041; respondents whose persona said "conservative" actually voted 20 Democratic / 15 Republican. Re-measured on the repaired artifact: the mode-decoded party field reaches 0.251 (§5.2), while a single draw at the production temperature carries

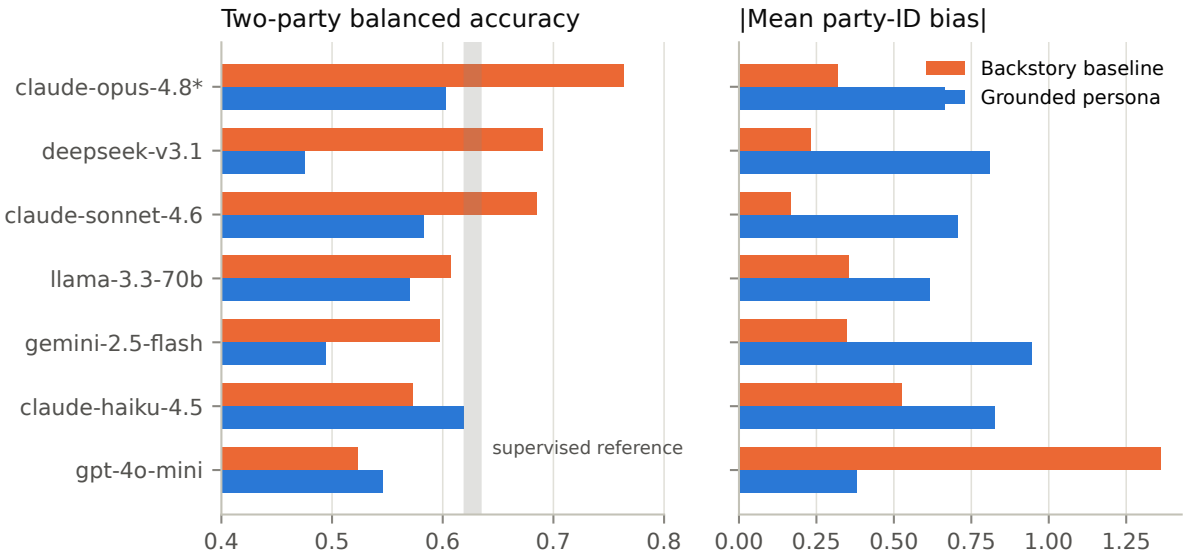

Figure 1: Cross-model study. **Left:** two-party balanced accuracy; the gray band is the supervised demographics-only reference (logistic regression to random forest). **Right:** absolute mean party-ID bias. The grounded arm (blue) helps only on `gpt-4o-mini`, whose vanilla baseline (orange) is pathologically Democratic-skewed. * N=60.

−0.045—nothing, by design (§5.3). This is not a defect of the sampler: ideology inferred from seven demographics is a deterministic-plus-noise function of X, and Proposition 1 applies. (The supervised reference gives the empirical scale: a well-specified predictor extracts about 0.63 balanced accuracy from these demographics on this sample.)

2. **The elicitor is substantially obedient, with model-dependent strength.** In the pilot, correlation between the injected ideology label and the predicted party was 0.893. Re-measured per model on the repaired cross-model run: obedience ranges from 0.56 (claude-haiku) to 0.81 (deepseek). Obedient enough that the label displaces much of the LLM's own demographic prior—which is strong (the baseline scored 24/24 on Black respondents in the pilot)—but not the constant the pilot suggested, and the least obedient models are the ones whose grounded arms sit closest to their baselines.

3. **Therefore arm accuracy tracks label accuracy.** An obedient reader of a near-random label cannot beat the reader's own priors. This is the empirical content Proposition 1 leaves open: whenever the model's $\hat{p}(Y \mid X)$ is decent, displacing it with a noisy compression of the same X is a strict loss; only when that prior is badly biased does the label's population-level correctness compensate. The 5-of-7 split in Table 1 is exactly this pattern, ordered by baseline bias.

Two implications follow. For evaluation: *no* single-draw, single-shot individual-prediction design can reward population-grounded persona generation, because the mechanism it adds (correct conditional *distributions*) is invisible at the individual draw. For deployment: whatever a persona system injects *will be obeyed*, so injected fields must be right at the level the simulation is used—which motivates Study 2's shift of both the evaluation and the engineering to the population level.

Why a single calibrated draw is a coin flip. Table 1 already runs on the repaired artifact, so the negative result is not an artifact of the decode defects. The repair makes the arithmetic visible. A single calibrated draw's party field matches the respondent's true vote 50.9% of the time—individually a coin flip *by design*, because the calibration of §5.3 restores the population's near-maximal within-cell heterogeneity as per-draw noise. Sampling one attribute value from a correctly calibrated conditional is the worst possible individual point estimate the conditional supports.

This sharpens the scope of the claim. Population fidelity is opposed to *sampled-draw* accuracy, not to individual prediction as such: the same calibrated conditional, read out as a probability rather than as a sampled character, remains a perfectly good probabilistic prediction and would score well on log loss or Brier score. What the grounded pipeline cannot do is beat the elicitor's own $\hat{p}(Y \mid X)$, by Proposition 1. What it should not be asked to do is answer with one draw when the question wants a distribution.

5 Study 2: population-level fidelity and steering

5.1 Re-scoped, preregistered hypotheses

Study 1's design predicted one draw per respondent at latent temperature 0 (a mode decode, not a draw from the joint), rendered with research-local templates, and never scored the fusion model's directly-decoded fields. The product mechanism is different on every axis: a user specifies an audience *cell*; the sampler draws K personas from the conditioned joint; the readout is a *distribution*. Argyle et al. (2023)'s own construct is distributional. Study 2 therefore locks two hypotheses, with falsification criteria, before any paid run:

> **H1′ (population calibration).** For demographic cells of the ANES 2020 voter population, the response distribution from eliciting K fusion-sampled personas per cell (product configuration and rendering) is closer—in total-variation distance, cross-cell share correlation, and within-cell dispersion fidelity—to the real survey-weighted conditional distribution of vote and party ID than K elicitations of the same base LLM given an identical cell-level backstory at matched temperature.
>
> **H2′ (steering dose-response).** When the audience specification extends beyond demographics to criteria the fusion posterior conditions on (preregistered knobs: ideology band, religious attendance), the induced vote/party distributions track the real ANES conditionals for the same subgroups—including monotone dose-response across knob levels—with lower error than the same base LLM given the identical criteria verbally.

Falsifiers: for H1′, the vanilla-K arm matching or beating on TV for either target, or the grounded arm showing materially worse dispersion (ratio outside $[0.75, 1.33]$); for H2′, no TV advantage on steered cells, or non-monotone dose-response where the truth is monotone.

5.2 Auditing the generative layer: two silent decode defects

Study 1 logged the fusion model's decoded `party_id` for every persona but never scored it. Scoring it first—the free arm—was this study's gating stage, and it immediately surfaced corruption: on the pilot sample the decoded party field read 9 Democratic / 87 Republican / 46 "other" / 5 literal `'8'` against a true 113/77 split. Root-causing found two compounding defects:

1. **Vocabulary pollution.** One training source (CES) declared a code scheme only for its 7-point party item; 3-point-only waves fell through to a generic label fallback that passed raw numeric codes into the training vocabulary—16,741 junk rows (codes for "other," "not sure," and "skipped") entering the model as if they were party labels.

2. **Categorical-as-index decoding.** The trainer and decoder treated categorical fields as scalar regression on the *alphabetically sorted category index*—meaningless for unordered domains. With vocabulary [`"4"`, `"5"`, `"8"`, `democrat`, `independent`, `other`, `republican`], the most-Democratic cells regressed *past* `democrat` into the junk codes: four of the five personas decoded as `'8'` were true Strong Democrats.

The fix closes the party vocabulary at three layers, redeclares `party_id` as an ordinal field (democrat < independent < republican—the one shape the scalar-decode architecture handles correctly), and retrains. Validation on the pilot protocol: individual-level correlation of the decoded party with true partisanship rises $0.041 \rightarrow 0.251$, and the decoded field alone—no LLM—predicts vote at 0.630 balanced accuracy, matching the supervised reference. *The grounding carried supervised-level signal all along; the decode was destroying it.* Every conclusion in Study 1 about "the grounding" was really about the grounding as corrupted by this defect; the mechanism dissection of §4.4 explains why fixing it cannot change Study 1's verdict (the individual ceiling binds either way), but Study 2's population claims are only testable after the repair.

We highlight the audit as a methods contribution in its own right: the defect was invisible in rendered persona text, invisible in end-to-end LLM metrics (which blur generative and elicitation error), and obvious the moment the tabular layer was scored against ground truth. **Score the generative layer directly; it is free.**

5.3 Post-hoc decoder calibration

With the semantic defect fixed, a free sweep of the fusion-only arm over all preregistered cells (K=200 per cell) quantified two residual gaps: a rightward marginal bias (unconditional draws 45% Republican / 31% Democratic vs. training-population 37%/48%)—the field heads are fit on survey-row latents but applied to ACS-backbone cell-mean latents, and nothing calibrated that transfer—and under-dispersion (within-cell SD ratio ∼0.87), because the decode added no idiosyncratic variance. Both were repaired post-hoc, without retraining:

1. **Cutpoint calibration.** Per-field decision cutpoints on the prediction axis, fitted by weighted quantile matching of the ACS-population prediction distribution (Monte Carlo over all 270,479 backbone cells at the operating latent temperature) to the survey-weighted training marginal. After fitting, unconditional marginals are essentially exact (party 44/39/18 vs. target 45/37/19; ideology within one point everywhere).

2. **Dispersion matching.** Per-field idiosyncratic decode noise σ, fitted by bisection so the sampler's within-cell SD of the decoded index matches the *true* within-cell SD in the training population over the same cell specification (party target 0.882 on the 0–2 axis $\rightarrow \sigma = 2.62$). Cutpoints are refit jointly with σ. The magnitude is itself a finding: the true conditional variance of politics given coarse demographics is close to maximal—an independent confirmation of Study 1's information bound—so the calibrated fusion layer's contribution is *conditional means and their calibration*, with honest noise supplying the heterogeneity.

Table 2: H1′ results (all 74 cells): survey-weighted mean TV distance to real cell conditionals (lower is better; bootstrap 95% CIs over cells) and party-ID dispersion ratio (1.0 is perfect; preregistered gate $[0.75, 1.33]$). The grounded arm beats the preregistered vanilla-K arm decisively; verbalized sampling ties it; the direct arm wins both static readouts (its party-7 number is from a dedicated distribution elicitation, temperature 0, two repeats pooled).

Base model	Arm	Vote TV [95% CI]	Party-7 TV	SD ratio
deepseek-chat-v3.1	A′ grounded personas	**0.112** [0.091, 0.133]	0.342	**0.81**
deepseek-chat-v3.1	B′ vanilla-K	0.368 [0.327, 0.403]	0.571	0.49
deepseek-chat-v3.1	V verbalized sampling	0.117 [0.105, 0.129]	**0.318**	**0.85**
deepseek-chat-v3.1	D direct readout	**0.041**	**0.132**	—
gpt-4o-mini	A′ grounded personas	0.155 [0.131, 0.179]	**0.365**	**0.92**
gpt-4o-mini	B′ vanilla-K	0.382 [0.320, 0.437]	0.682	0.24
gpt-4o-mini	V verbalized sampling	**0.129** [0.121, 0.140]	0.387	**0.85**
gpt-4o-mini	D direct readout	**0.045**	**0.192**	—

Appendix G states both procedures formally, with pseudo-code, the progression figure, and an analysis of what they do and do not guarantee about joint dependence. At the gating stage the progression is: total-variation distance 0.417 → 0.236 → 0.166 → 0.130, dispersion ratio 0.82 → 0.994, junk mass 30% → 0. The same sweep found that higher latent temperature monotonically improves both TV and dispersion; the product default was raised from 0.1 to 1.0 accordingly. The feared tension between added noise and cell-mean fidelity did not materialize: real per-cell distributions are themselves wide, so honest dispersion *improves* TV.

5.4 Population calibration (H1′)

Design. The unit of analysis is the cell: all levels of the six one-way demographic marginals plus two preregistered two-way crossings (education × race, age band × region), floor of 30 effective respondents—74 cells. Four preregistered arms per cell: **F** (fusion-tabular: K=200 decoded draws, no LLM); **A′** (the product path: K=50 fusion draws → product rendering → one elicitation each); **B′** (vanilla-K: the cell's attributes as a backstory, elicited K=50 times at the same temperature); **D** (direct: one call asking the base model to state the subgroup's distribution). A fifth, post-hoc arm probes whether the vanilla failure is prompting-specific: **V** (verbalized sampling, Zhang et al., 2025): each call asks for *five distinct* plausible responses from different voters matching the description, with verbalized probabilities; 20 calls per cell pool 100 probability-weighted candidates—the same candidate budget as B′ at a fifth of the calls. Two base models (`deepseek-chat-v3.1`, the hardest test as Study 1's best-calibrated baseline, and `gpt-4o-mini` for continuity), elicitation temperature 0.7 everywhere, two repeats pooled, 29,896 calls in the main design, every prompt logged. Ground truth: survey-weighted ANES conditionals over the full 6,222-voter population.

A correction. An earlier execution of this design silently failed 94.7% of its calls and scored only the surviving 4 of 74 cells while reporting completion; all Stage-1 numbers here come from a clean, fully-successful rerun (zero failed calls). Appendix I documents the failure, how it was caught, and the harness changes—reported because it is the run-log corollary of this paper's central methodological claim: *score coverage, not completion counts.*

Result: H1′ supported on both models (Table 2, Figure 2). Against the preregistered vanilla-K arm, the grounded arm's vote TV is 3.3× better on deepseek and 2.5× better on gpt-4o-mini, with non-overlapping bootstrap CIs; party-7 TV falls by a third to a half; and the

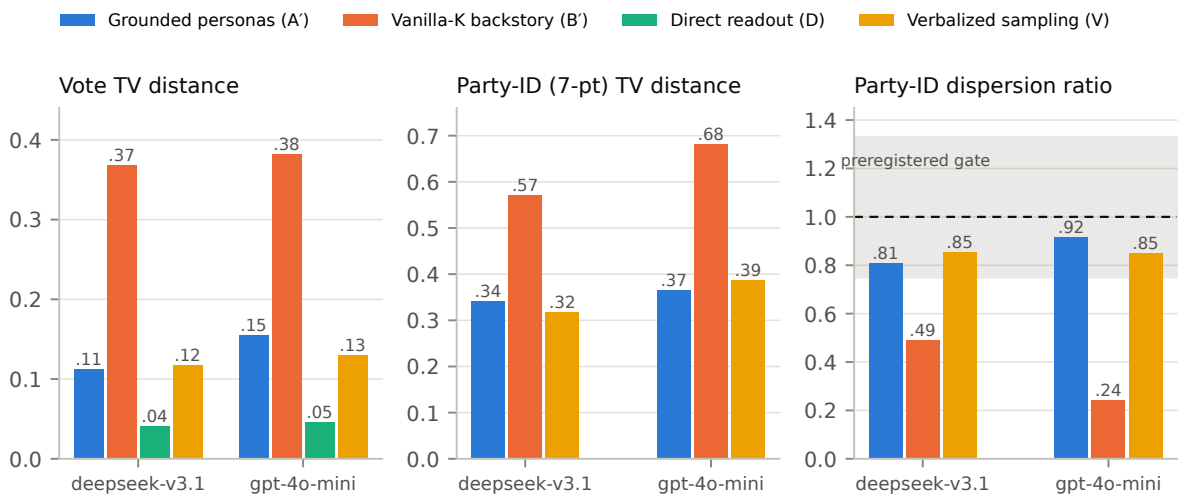

Figure 2: H1′: population calibration over 74 ANES cells. Grounded K-persona simulation (blue) cuts TV distance to roughly a third of vanilla-K prompting's (orange) on both targets and both models, and sits inside the preregistered dispersion gate while vanilla-K collapses variance. The verbalized-sampling arm (yellow) ties the grounded arm and restores dispersion; the direct readout (green) is the strongest static baseline on both targets.

grounded dispersion ratio (0.81 / 0.92) sits inside the preregistered gate while vanilla-K collapses to 0.49 and 0.24: the Bisbee et al. (2024) variance collapse, isolated. Four observations:

1. **The failure of vanilla-K is dispersion, not just level.** K elicitations of one backstory at temperature 0.7 are not a population; they are one modal answer with sampling noise. Consistently, vanilla-K's error is *flat in* K (Appendix J): a collapsed distribution cannot buy fidelity with more samples, while the grounded arm's error keeps falling as personas accumulate.

2. **The product pipeline preserves the generative layer's calibration.** Grounded vote TV (0.112/0.155) brackets the fusion-only arm's ∼0.12–0.13: rendering personas to text and passing them through an LLM does not destroy the calibrated distributions. This was the operative engineering question for the product.

3. **The direct arm wins static readouts outright.** Asked for a subgroup's vote shares, both models state distributions closer to the ANES conditionals (TV 0.041/0.045) than any simulation produces; asked for the full 7-point party distribution, the same holds (TV 0.132/0.192 vs. the best simulation's 0.318). Modern LLMs have well-calibrated conditional knowledge of American political distributions, not just their modes.

4. **Verbalized sampling ties the grounded arm—and reframes the vanilla failure.** The V arm matches A′ on population distributions (deepseek: CI-overlapping tie; gpt-4o-mini: V significantly better on vote TV) with dispersion restored to 0.85, at a fifth of the calls. Vanilla-K's collapse is therefore an artifact of the *one-character-repeated* prompting pattern, not an inherent limit of unguided LLMs: asked to verbalize a distribution over different people, the base model's conditional knowledge is well-calibrated and appropriately dispersed. What V does not provide is what no distribution readout provides: coherent individual personas whose many attributes cohere jointly, calibration traceable to named surveys rather

Table 3: H2′ results over 27 steered cells: TV distances plus dose-response quality per knob (Pearson correlation and MAE of predicted vs. true Republican two-party share across knob levels, pooled over contexts).

				Ideology		Attendance	
Model	Arm	Vote TV	Party-7 TV	corr	MAE	corr	MAE
deepseek	A′ grounded	**0.111**	**0.375**	**0.963**	**0.112**	0.538	0.099
deepseek	B′ vanilla	0.244	0.489	0.937	0.149	0.729	0.310
deepseek	D direct	0.087	—	0.956	0.121	**0.803**	**0.080**
gpt-4o-mini	A′ grounded	**0.148**	**0.420**	**0.949**	**0.128**	−0.009	0.162
gpt-4o-mini	B′ vanilla	0.296	0.630	0.930	0.158	0.515	0.436
gpt-4o-mini	D direct	0.149	—	0.920	0.150	**0.588**	**0.136**

than to pretraining priors, and agents that can enter downstream interaction. The static-distribution franchise belongs to prompting (D for one-way marginals, V generally); the grounded arm's case rests on the rest.

5.5 Steering (H2′): coupling-bounded dose-response

Design. Two preregistered knobs the fusion posterior conditions on and ANES independently measures: ideology band (five levels, ANES 7-point self-placement collapsed) and religious attendance (four levels). Each knob sweeps its levels within three contexts (national; region = South; education = bachelor's), mass floor 30—27 cells, same arms and models, 10,908 calls. The vanilla arm receives the identical criteria as plain first-person sentences. Conditioning on ideology to read the *vote* distribution is not leakage; it is the steering claim itself, and it is not trivial—the correct conditional differs sharply by context.

Result: supported on deepseek; not fully supported on gpt-4o-mini (Table 3, Figure 3). The grounded arm's TV is roughly half the vanilla arm's everywhere (first criterion passes on both models), and its dose-response is monotone on deepseek for both knobs—but the attendance curve on gpt-4o-mini is non-monotone ($0.28 \to 0.31 \to 0.28 \to 0.37$ against a true $0.35 \to 0.47 \to 0.50 \to 0.59$), failing the second criterion. The national-context curves carry the story:

- **Verbal steering produces caricature.** The vanilla arm maps "never attends" to near-certain Democrat and "attends weekly" to *certain* Republican (deepseek: $0.07 \to 0.00 \to 0.55 \to 1.00$; gpt-4o-mini: $0 \to 0 \to 0 \to 1.00$), and renders ideology as a step function. Direction right, levels absurd—stereotype amplification (Cheng et al., 2023) measured on a real criterion. Its decent correlations and terrible MAEs (0.31–0.44) are the signature.

- **Grounded steering is bounded, and the bound is measurable.** The knob is pinned on the persona exactly (the evidence clamp), but everything *else* in the profile moves only as far as the posterior propagates the knob. Ideology steers vote nearly perfectly ($r = 0.949$–0.963, best MAE); attendance under-propagates, flat on one model (deepseek $0.286 \to 0.460$ against a true $0.345 \to 0.594$). The fusion-only audit of §5.6 decomposes *why*, and the decomposition is not the obvious one: the posterior-propagation channel is heavily attenuated for *both* knobs, and the ideology knob's success is carried by the rendered label itself.

- **Both simulation arms polarize the middle.** At the moderate ideology level, both under-shoot the true Republican share badly (grounded 0.071 vs. true 0.329 nationally);

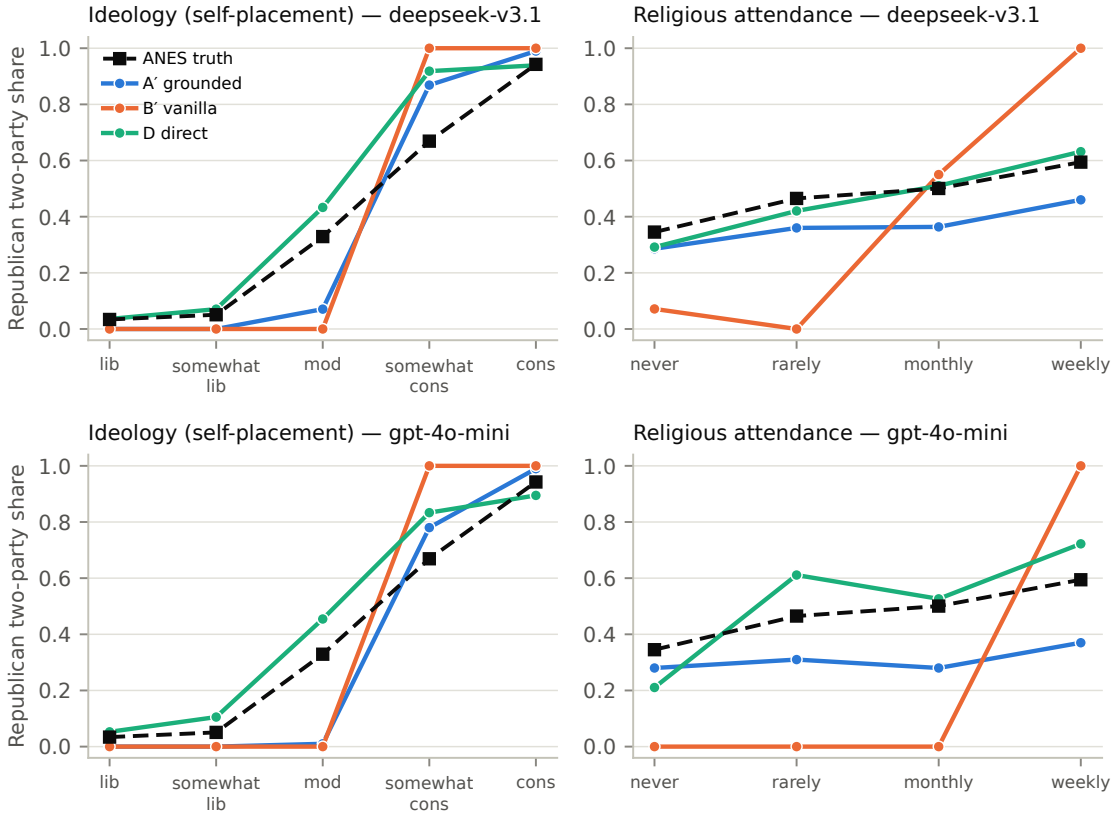

Figure 3: Steering dose-response in the national context (truth gradient dashed black). The vanilla arm (orange) is a caricature machine: step functions from certain-Democrat to certain-Republican. The grounded arm (blue) has sane levels but an under-steered attendance gradient—flat on `gpt-4o-mini`—because attendance→politics propagates through a weakly learned cross-field coupling. The direct readout (green) tracks the gradient best.

the direct arm does not (0.433). Moderates are the hardest population for in-character simulation.

- **The direct arm remains the best static readout** on attendance dose-response—the Stage-1 pattern persists into steering for criteria the base model has world knowledge about.

Verbalized sampling on steered cells. Over the same 27 cells, the V arm extends its Stage-1 showing: steered vote TV 0.141 on both models, steered party-7 TV 0.285/0.292 (better than A′'s 0.375/0.420), ideology dose-response $r = 0.949/0.969$, and attendance $r = 0.695$ on *both* models with MAE 0.08–0.12—no caricature, and better attendance tracking than the grounded arm's coupling-bounded 0.538/−0.009. In the two-channel language of §5.6, verbalized sampling is a better *label-channel* instrument than either B′ or A′: rather than sampling one obedient character per call, it asks the model to verbalize its conditional distribution directly, sidestepping both obedience caricature and the fusion layer's weak posterior propagation. The steering franchise for outcomes the base model has conditional knowledge about therefore also belongs to prompting; grounded steering remains necessary exactly where that knowledge runs out—un-memorized intersections, novel attributes, and knobs that must move a whole coherent persona rather than a stated share.

Table 4: Fusion-only coupling audit over the 27 steering cells (K=200 decoded draws per cell, calibrated artifact, latent temperature 1.0). Outcome: decoded party identification, scored against the ANES party conditional (pid7 collapsed) for the same cells. Both knobs' posterior propagation is heavily attenuated; attendance is uncoupled outright. (Attendance MAE is small only because the true gradient is modest and the flat fusion readout sits near its middle; the correlation is the coupling measure.)

Knob	Dose-response r	Rep.-share MAE	Mean party TV
Ideology (5 levels × 3 contexts)	0.61	0.32	0.37
Religious attendance (4 × 3)	−0.11	0.11	0.16

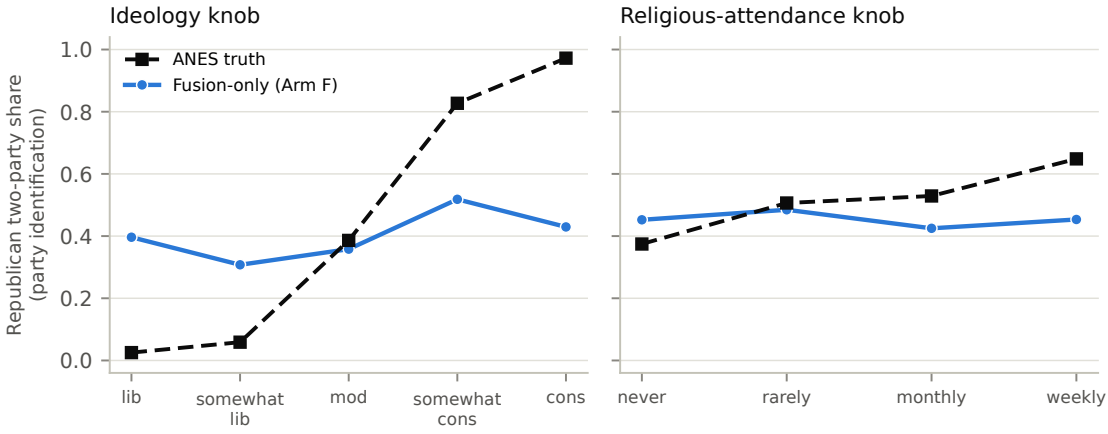

Figure 4: The audit's national-context curves: fusion-only decoded party share (blue) against the ANES party conditional (dashed). The posterior barely moves the decoded politics under either knob—the ideology gradient spans 0.31–0.52 against a true 0.03–0.97. Stage 2's near-perfect LLM-level ideology steering therefore cannot be coming from this channel.

5.6 The coupling audit

The steering bound is measurable in advance, for free. **The audit:** for a candidate knob k and outcome field Y, (1) enumerate the steering cells (knob level × context, mass floor as in §5.5); (2) draw K personas per cell from the fusion posterior alone—no LLM—and read the *decoded* Y; (3) score the induced dose-response of Y across knob levels against a reference conditional (survey ground truth where available, else the training conditional): Pearson r across level–context points and MAE on the steered quantity. A flat or non-monotone fusion-only gradient predicts that LLM-level steering of everything the knob does not literally name will be attenuated. Cost: 27 × 200 persona draws, a few minutes of CPU.

The result (Table 4, Figure 4) is sharper than a naive coupling story, in an instructive way. The audit *predicts the attendance failure cleanly*: no coupling at the fusion layer ($r = -0.11$), flat-to-non-monotone steering at the LLM level. But it also shows the ideology knob's posterior propagation is itself heavily attenuated ($r = 0.61$, with the decoded gradient spanning 0.31–0.52 against a true 0.03–0.97)—yet Stage 2 measured near-perfect LLM-level ideology steering. The resolution is that steering flows through *two channels*:

1. **The label channel.** The clamped knob value is rendered into the persona text, and the elicitor obeys rendered labels (Study 1's $r = 0.893$ obedience, working productively here). This channel is strong exactly when the knob is semantically proximal to the outcome—"I'd

Table 5: Same-data statistical baselines against the fusion-only arm, 74 cells, party distribution, survey-weighted TV against ANES (lower is better) and within-cell dispersion ratio (1.0 ideal). TV for M0–M2 is computed on their exact distributions; the fusion arm's is from K=200 draws—see the symmetrization note in the text before comparing.

Population model (same training data)	Party TV	Dispersion ratio
M0 pooled training marginal (no conditioning)	0.138	0.971
M1 weighted empirical conditional	**0.103**	0.952
M2 multinomial logit, marginalized	0.129	0.929
Arm F: fusion latent model (calibrated)	0.133	**0.986**

say I'm conservative" reads directly onto vote—and it is the channel the vanilla arm uses, in caricatured form. The vanilla-K sweep is in fact its clean measurement: B′ ideology $r = 0.93$–0.94, attendance $r = 0.52$–0.73, both with absurd levels.

2. **The coupling channel.** The posterior propagates the knob into the *rest* of the profile—most importantly the sampled ideology field that Study 1 showed anchors the elicitor. This is the channel the audit measures, and it is weak for both knobs in the current artifact.

The grounded arm composes the two: for the ideology knob, the label channel drives the elicitor while the grounded profile supplies sane levels (A′ $r = 0.95$–0.96, best MAE); for the attendance knob, the label channel is weakly outcome-proximal, the coupling channel leaves the profile's ideology anchor flat, and the anchor wins—the under-steered gradients of Table 3. This also explains why population calibration (H1′) succeeds while knob propagation is weak: demographic conditioning operates through the ACS backbone reweighting, over directly observed joint survey rows; non-demographic knobs operate through support reweighting plus a small latent nudge, and the eight-dimensional latent bottleneck attenuates them. Strengthening that pathway—more latent capacity, cross-source anchors, joint knob–politics sources—is the concrete fusion-v2 roadmap, and the audit is the free instrument that will tell us when it has worked.

6 What the advantage is, and what it is not

Sections 5.4 and 5.5 establish that grounded sampling beats repeated prompting and that prompting beats it back on static readouts. Two further comparisons locate what remains.

6.1 Same-data statistical baselines

The grounded generator is fit on 4.87M survey observations; the arms it has been compared against so far are prompting arms with no access to that data. The fair question is how much of its population fidelity is available from the same training data with textbook machinery and no latent model. We built three such baselines over the identical 74 cells, scored with the identical instrument (Appendix H): **M0**, the pooled training marginal, identical for every cell; **M1**, the survey-weighted empirical conditional with hierarchical backoff when a cell is thin—what a survey analyst would do; and **M2**, a weighted multinomial logit on one-hot demographics, marginalized over the attributes each cell leaves free.

One instrument detail matters before reading the comparison, because it initially misled us: the statistical baselines return exact distributions, while the fusion arm's score comes from K=200

Table 6: Decomposing the grounded arm, 74 cells, same elicitor and K as Table 2. **L** supplies the cell backstory plus the draw's ideology label only; **S** supplies the full persona with the political stance and psychographic block independently permuted across draws (marginals preserved, joint destroyed). Most of the advantage over B′ is already present in L.

Base model	Arm	Vote TV	Party-7 TV	Dispersion
deepseek-chat-v3.1	B′ vanilla-K (no grounding)	0.368	0.571	0.49
deepseek-chat-v3.1	L label injector	0.126	0.409	0.85
deepseek-chat-v3.1	S shuffled joint	0.122	0.356	0.79
deepseek-chat-v3.1	A′ full grounded persona	**0.112**	**0.342**	0.81
gpt-4o-mini	B′ vanilla-K (no grounding)	0.382	0.682	0.24
gpt-4o-mini	L label injector	**0.131**	0.400	1.00
gpt-4o-mini	S shuffled joint	0.157	0.377	0.91
gpt-4o-mini	A′ full grounded persona	0.155	**0.365**	0.92

sampled draws, which carry pure sampling noise (≈0.01–0.02 TV at this K; Appendix J). Scored asymmetrically, the lookup appears to beat the generator by 0.030. Symmetrized—sampling the lookup's distributions at the same K=200—M1 scores 0.113 against the fusion arm's 0.118–0.133 (that range is run-to-run variance between K=200 instruments). The honest verdict: *a weighted lookup on the generator's own training data matches the generator* at reproducing coarse-cell conditionals. The latent model neither beats the table it approximates nor loses to it; it earns nothing on this task. Two qualifications keep this in proportion, and neither rescues the generator here. First, the fusion arm's dispersion is closest to truth (0.986), but that is fitted by construction (§5.3), so it is not independent evidence. Second, and substantively: M0–M2 return a distribution for a cell and nothing else. They cannot emit an individual with jointly coherent attributes, cannot condition on anything outside the demographic table, have nothing to say about a cell with no training support, and cannot be rendered into a persona that enters a conversation. Population-distribution readout on well-supported demographic cells is not a task that justifies a persona generator—the same verdict the prompting arms deliver, now from the data side.

6.2 Is it just a calibrated label injector?

Study 1's renderer sweep found that one grounded field, ideology, carries nearly all political behavior (§4.2). If that holds at the population level, the grounded arm is a calibrated ideology-label injector and "joint structure" is interpretation rather than mechanism. Two ablations over the same 74 cells, same elicitor, same K and temperature, decide it: **L** (label injector) gives the model the cell-level demographic backstory used by B′ plus the k-th draw's ideology sentence and nothing else, isolating the calibrated label distribution; **S** (shuffled joint) renders the full product persona but permutes the political stance and the psychographic block independently across the K draws within each cell, so every field's within-cell marginal is preserved exactly while the cross-field joint is destroyed.

The answer is mostly yes, and we report it as such. On *vote share*, the label injector is statistically indistinguishable from the full persona on deepseek (0.126 vs. 0.112, inside the design-aware interval of Appendix J) and *better* than it on gpt-4o-mini (0.131 vs. 0.155). One calibrated ideology sentence appended to the vanilla backstory recovers essentially all of the grounded arm's three-category advantage: TV falls from B′'s 0.368/0.382 to 0.126/0.131, and dispersion recovers from 0.49/0.24 to 0.85/1.00. Whatever the persona pipeline is doing for a coarse binary outcome,

a single well-calibrated label does too.

On the finer *seven-point* distribution, adding psychographic content at all—even incoherently, arm S—improves TV over the label alone by 0.052 (deepseek) and 0.023 (gpt-4o-mini), both separated from zero in a paired per-cell bootstrap. Making that content jointly *coherent* rather than shuffled shows a further +0.015 and +0.011—but those intervals span zero on both models ($[-0.004, +0.032]$ and $[-0.000, +0.021]$). The vote-level picture is mixed in both directions: on deepseek the full persona beats both L and S by small, separated margins; on gpt-4o-mini the label injector beats the full persona, also separated. So coherence, at this design's resolution, is not distinguishable from no effect.

We draw three conclusions, two of which cut against our own earlier framing. First, the population-level result of §5.4 is substantially a *calibrated-label* result: the fusion layer's contribution to these distributions is dominated by getting one marginal right, which is precisely what §5.3's cutpoint calibration was built to do. Second, per-persona joint coherence—the property the correlated-noise calibration restores and the joint-dependence audit certifies—has no detectable effect on these readouts; its value, if any, lies in uses these readouts cannot see, such as how a coherent persona behaves across many items or in conversation, which we have not measured. Third, for anyone choosing a tool: a practitioner who needs a coarse subgroup distribution should inject a calibrated label (or use a lookup, §6.1) rather than run a persona pipeline. The case for the generator rests entirely on what no distribution readout can provide.

7 Discussion

Match the evaluation to the level of the claim. The same system, on the same held-out survey, is refuted at one level of analysis and supported at another. Study 1's negative result is real and, by Proposition 1, not fixable by richer sampling: population-drawn attributes carry no respondent-specific information. Study 2's positive result is equally real at the level where the product claim lives. Much of the current literature evaluates persona methods at whichever level is convenient; the verdict can flip entirely with the level. Claims about simulated *populations* should be scored on distributions, dispersion, and steerable conditionals; claims about simulated *individuals* need individually-measured grounding of the kind Park et al. (2024) collect, not population draws.

Whatever you inject will be obeyed—so audit what you inject. The elicitor's near-perfect obedience to injected attributes is the pivot of both studies: it is why a noisy label destroys individual prediction, why a calibrated label distribution builds a good population, and why steering works through the rendered label more than through the posterior. Injection is a bet that the injected value beats what it displaces, at the level of the use. Two audits price that bet before deployment, both free and both specified here: score the generative layer's decoded fields directly against ground truth (§5.2), and score each steering knob's propagation with no LLM in the loop (§5.6). Persona systems have a tabular layer under the prose; treating it as unscoreable internal detail is how corruption persists—in our case for months, invisible in rendered text and in end-to-end metrics. The same discipline applied to our own run logs caught a silently failed experiment whose aggregates looked complete (Appendix I).

Dispersion is a first-class metric. Level-based metrics alone hide both halves of the variance story: the vanilla collapse to 0.24–0.49 of real within-cell spread, and the fact that a different

prompting pattern restores it entirely. We recommend the dispersion ratio, with a preregistered acceptance band, as standard practice for silicon-sample evaluation (following Bisbee et al., 2024).

When to simulate rather than ask. Our prompting and statistical arms are a standing challenge to the simulation literature. For one-way demographic conditionals of well-surveyed populations, asking the model directly is cheaper and more accurate than any persona pipeline; for full response distributions, verbalized sampling matches grounded simulation at a fifth of the cost; and for the same cells, a weighted lookup on the generator's own training data matches the generator (§6). Static distribution readout, as a task, does not need a persona simulator. And §6.2 narrows the claim further from inside: on coarse outcomes a single calibrated label recovers the grounded arm's advantage, so even the population result is mostly a calibration result. What the alternatives do not provide is a population of *individuals*: per-persona draws whose attributes cohere jointly (certified by the audits of Appendix G, though undetectable in these readouts), calibration traceable to named surveys rather than implicit in pretraining, conditionals outside the model's world knowledge, and agents that can enter downstream *interaction*—conversation, deliberation, group dynamics—where there is no single question whose answer is the deliverable. This paper establishes the middle two and cannot speak to the last, which is where the case for persona simulation now rests and which the companion study on deliberation dynamics takes up.

Engineering calibration versus learning it. Our decoder repair is post-hoc and model-internal; Cao et al. (2025) reach the same target by fine-tuning the LLM against first-token distributional divergence. Fine-tuning bakes calibration into the language model (general across elicitation formats, but opaque and retraining-bound); generative-layer calibration is transparent and cheap to refit but must be maintained per field and operating point. Given §5.6, we expect them to be complementary—persona layer and elicitor layer—and a direct comparison on shared cells is the natural next experiment.

8 Limitations

Scale and models. The pilot ran at N=200; cross-model runs at N=150/N=60; population stages on two base models. Effect directions were consistent across every model and stage tested, and the population-stage margins are CI-separated (Appendix J), but the small-N comparisons (notably the interview-renderer effect and per-model cross-model differences under $\sim$0.08) are directional rather than conclusive, and H2′'s split verdict already demonstrates base-model dependence. **Backend nondeterminism.** Temperature-0 completions from commercial APIs drifted between identical runs by margins comparable to small effects; we pooled repeats everywhere after discovering this, but it bounds the resolution of all LLM-arm comparisons. **One criterion, one country, one cycle.** All results concern U.S. politics, ANES 2020, self-reported voters, and two political targets; the sample skews Democratic (59% two-party), which raw accuracy rewards—we report balanced accuracy for this reason. Generalization to other domains, countries, and less-surveyed populations is untested, and the direct-readout arm's strength should *shrink* where pretraining coverage is thinner. **The advantage is mostly calibration.** §6.2 shows a single calibrated label recovers most of the population-level advantage on coarse outcomes, and §6.1 shows a weighted lookup on the training data beats the generative layer at reproducing those conditionals. The residual attributable to joint persona structure is not separable from zero at this design's resolution, and was probed on one attribute pair in one domain. **Coarse mappings.** Attendance ground truth uses ANES self-reports with a coarse level mapping; the

direct arm's 7-point score comes from a separate distribution elicitation rather than the Stage-1 protocol. **Steering coverage.** Two knobs, three contexts, at the LLM level. The fusion-only audit (§5.6) covers both knobs and correctly predicts the attendance failure, but the two-channel account has been *validated* against LLM behavior only on this pair; candidate knobs the audit should triage next—Big Five bands, moral-foundation emphasis, generalized trust—await both the free audit and an LLM-level spot check before the contract can be called general. **Preregistration was internal.** Hypotheses, metrics, cells, and decision rules were locked in version-controlled documents before the corresponding runs, with deviations logged, but not registered with an external service. **Author interest.** The author develops the system under test; the design mitigations are the locked decision rules, the within-run supervised and direct-readout baselines, and the fact that the paper's first study is a self-refutation.

9 Ethics and broader impact

Steering as a dual-use capability. A system that produces *calibrated* synthetic publics, steerable by ideology or religiosity, is useful for the same reason it is abusable: it can be pointed at message testing, persuasion optimization, and microtargeting rehearsal against subpopulations who never consented to being simulated. Two properties of our results bear directly on the risk. First, the steering that works is bounded and legible: the per-knob audits of §5.6 are cheap to run and to publish, and we argue vendors of persona systems should ship them as disclosures—a steering claim without its audit is marketing. Second, the failure modes are not neutral: the vanilla arm's caricature machine (never-attenders $\rightarrow$ certain Democrat, weekly attenders $\rightarrow$ certain Republican) is exactly the stereotype-amplification pattern that harms the groups being simulated (Cheng et al., 2023), and miscalibrated conditioning concentrates distortion in underrepresented subgroups (Taday Morocho et al., 2026). Calibration work of the kind reported here reduces that harm channel but also makes the tool more credible; we think the balance favors publishing the calibration methodology together with its audits, so that claims about synthetic publics can be checked rather than trusted.

Synthetic respondents do not replace human data. Every result here is disciplined by a real probability survey; the direct-readout caveat and the individual-level negative result are both arguments *against* treating LLM personas as substitutes for measurement (cf. Bisbee et al., 2024; Zhao et al., 2025). The appropriate uses are diagnostic and exploratory: piloting instruments, stress-testing designs, and simulating interaction settings that cannot be directly asked.

Data and privacy. All grounding sources are public-use, licensed survey and administrative microdata; the generative model's backbone is a cell lattice over coarse demographic categories, and sampled personas are draws from modeled conditional distributions, not records of real individuals. ANES is used under its terms as a held-out evaluation criterion and never enters training.

The obedience warning as a safety property. The elicitor's near-perfect obedience to injected attributes (§4.4) means persona systems silently launder whatever their generative layer emits—defects included, as our decode audit showed. Scoring the generative layer directly is therefore not only a fidelity practice but a safety one.

10 Reproducibility and availability

Protocol reproducibility. All harness code, run configurations, prompts, raw completions, and per-run summaries are retained under version control, keyed by the run identifiers cited throughout. Seeds are deterministic functions of respondent/cell identifiers; fusion artifacts are pinned by name per study; figures are generated directly from the committed run summaries. The ANES 2020 Time Series Study is available from `electionstudies.org`; it is loaded into the research harness only and never enters the fusion model's training inputs. Appendices A–J specify the prompts, cell definitions, renderer variants, model equations, calibration procedures, and audit protocols at a level intended to support independent reimplementation of every experiment against a reader's own persona system.

Code and artifact availability. The persona system under test is commercial software, and its code and trained artifacts are not publicly released. Access is available to partners and clients under agreement, and we are exploring a hosted evaluation endpoint through which third parties could run the audits reported here (the fusion-only distribution scoring, coupling audit, and joint-dependence audit) against the deployed artifact without code access. The evaluation protocol itself carries no such restriction: ground truth is a public survey, the baselines (demographic backstories, direct readout, verbalized sampling) require nothing but an LLM API, and the audits are specified in full in the appendices.

Acknowledgments

The experimental harness, run orchestration, and drafting were carried out with LLM assistance under the author's direction; all locked designs, decision rules, and interpretations are the author's. Three AI-generated review passes from the Stanford Agentic Reviewer (`paperreview.ai`) informed successive drafts; the related-work coverage, the fusion-model and calibration appendices, the executed coupling and joint-dependence audits, the sensitivity analyses, the verbalized-sampling baseline, and the ethics section respond to their comments.

References

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32 (4):401–416, 2024. doi: 10.1017/pan.2024.5.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3141–3154, 2025. arXiv:2502.07068.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. arXiv:2310.11501.

Ziyan Cui, Ning Li, and Huaikang Zhou. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, 5 (8):627–634, 2025. doi: 10.1038/s43588-025-00840-7.

Ricardo Domínguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2306.07951.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2402.10811.

Minwoo Kang, Suhong Moon, Seung Hyeong Lee, Ayush Raj, Joseph Suh, David M. Chan, and John Canny. Deep binding of language model virtual personas: A study on approximating political partisan misperceptions. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2504.11673.

Mao Li and Frederick G. Conrad. Persona-based simulation of human opinion at population scale, 2026. arXiv:2603.27056.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2025.findings–emnlp.1261, 2025. arXiv:2507.16076.

Fernando Miranda and Pedro Paulo Balbi. Simulating public opinion: Comparing distributional and individual-level predictions from LLMs and random forests. *Entropy*, 27(9):923, 2025. doi: 10.3390/e27090923.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative agent simulations of 1,000 people, 2024. arXiv:2411.10109v1.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. arXiv:2303.17548.

Erika Elizabeth Taday Morocho, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. Assessing the reliability of persona-conditioned LLMs as synthetic survey respondents, 2026. arXiv:2602.18462.

Chenxiao Yu, Jinyi Ye, Yuangang Li, Zheng Li, Emilio Ferrara, Xiyang Hu, and Yue Zhao. A large-scale simulation on large language models for decision-making in political science, 2024. arXiv:2412.15291.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. Verbalized sampling: How to mitigate mode collapse and unlock LLM diversity, 2025. arXiv:2510.01171.

Jianpeng Zhao, Chenyu Yuan, Weiming Luo, Haoling Xie, Guangwei Zhang, Steven Jige Quan, Zixuan Yuan, Pengyang Wang, and Denghui Zhang. Large language models as virtual survey respondents: Evaluating sociodemographic response generation, 2025. arXiv:2509.06337.

A Prompt templates

All arms share one elicitor. System prompt:

```
You are role-playing as a specific American voter answering a short survey in
November 2020. Stay fully in character based only on the description provided.
Answer as this person would, even when unsure. Respond with ONLY a single
JSON object and nothing else.
```

Question (appended to every persona text):

```
In the 2020 U.S. presidential election, who did you vote for?
Answer with "vote": one of "biden", "trump", "other".
Then, which best describes your party identification?
"party_id": an integer 1-7 where 1=Strong Democrat, …, 7=Strong Republican.
Return exactly: {"vote": "...", "party_id": N}
```

Study 1 baseline arm (respondent-level backstory). Example:

```
I am a white woman in my late 50s/early 60s. I live in CA. I have a graduate
degree. My household income is over $150,000.
```

Study 1 grounded arm (interview rendering, validated variant). The persona is a fusion sample; only demographics and fusion-inferred fields appear—never the scored targets:

```
The following is an interview with one specific American, in late 2020.
INTERVIEWER: To start, tell me a little about yourself.
ME: I am a white woman in my late 50s/early 60s. I live in CA. …
INTERVIEWER: How would you describe yourself politically?
ME: I'd say I'm liberal.
INTERVIEWER: How would people who know you describe you, and what matters most
to you?
ME: They'd say I'm open to new experiences, organized, warm. What matters most
to me is fairness and loyalty to my group. I attend religious services weekly.
```

Study 2 grounded arm (product rendering). Example output of the product rendering path (`to_first_person_intro`) for one sampled persona:

```
INTERVIEWER: Tell me a little about yourself.
ME: I'm Maya Reed, 33, Female, and I work as Service worker in Management
Of Companies. My education is professional.
INTERVIEWER: And politically?
ME: I'd say I'm somewhat liberal.
INTERVIEWER: What kind of person are you, and what matters to you?
ME: by nature I'm curious, cooperative; what I care about most is high
generalized trust, weekly religious attendance, Fairness moral emphasis,
high local environmental burden.
ME: For background, I live in MI, and my household income is
$35,000-$75,000.
```

Study 2 vanilla-K arm (cell-level backstory). The cell's attributes as first-person sentences (for cell `sex=female`: "I am a woman."), elicited K times at temperature 0.7. Steered cells append the knob criterion verbatim (e.g., "I'd say I'm somewhat liberal." / "I attend religious services weekly.").

Study 2 direct arm. One call per cell:

```
Consider American adults who voted in the 2020 U.S. presidential election
with these characteristics --- sex: female.
Estimate: (1) the share who voted for Biden, Trump, or someone else; (2) the
mean 7-point party identification (1=Strong Democrat … 7=Strong Republican).
Return exactly: {"biden": 0.xx, "trump": 0.xx, "other": 0.xx, "party7_mean":
x.x}
```

B Preregistered cells

Population stage (74 cells): all levels of six one-way marginals—sex (2), race/ethnicity (6), age band (7), education (5), income band (4), census region (4)—plus the two-way crossings education $\times$ race and age band $\times$ region, subject to a floor of 30 effective (weighted) respondents per cell; sub-floor cells are dropped and reported. Steering stage (27 cells): ideology band $\in$ {liberal, somewhat liberal, moderate, somewhat conservative, conservative} and religious attendance $\in$ {never, rarely, monthly, weekly}, each crossed with three contexts (national; region = South; education = bachelor's), same floor. ANES ground-truth codings for the knobs (`V201200` ideology self-placement; `V201453` attendance) were verified against the codebook, with the 7-point-to-5-band ideology mapping recorded in the harness.

C Extended related work

Silicon sampling and algorithmic fidelity. Argyle et al. (2023) condition GPT-3 on demographic backstories from ANES respondents and show fine-grained, demographically-correlated correspondence between silicon and human samples on vote choice and related items. Park et al. (2024) interview 1,052 individuals for two hours each and build generative agents that answer held-out General Social Survey items at $\sim$85% of each person's own test–retest ceiling, versus $\sim$0.74 of ceiling for demographic-only agents; Kang et al. (2025) report large fidelity gains from interview-transcript "deep binding" backstories. These lines show that *rich individual-specific* conditioning helps. Our Study 1 asks a different question: whether *population-sampled* depth—plausible traits drawn from the joint distribution given demographics, not measured from the individual—helps. It does not, and §4.4 explains why it cannot.

Critiques of LLM survey simulation. Santurkar et al. (2023) document that instruction-tuned models default to liberal-leaning, educated response profiles; we measure exactly this skew as a -1.0-point party-ID bias in one widely-used model's vanilla baseline. Bisbee et al. (2024) show synthetic samples collapse variance and distort correlations; our dispersion-ratio metric operationalizes their warning and catches the collapse in the vanilla-K arm. Domínguez-Olmedo et al. (2024) show answer-order and label artifacts drive much apparent alignment; Cheng et al. (2023) characterize caricature in persona simulations, which our steering experiment measures as step-function dose-response; Hu and Collier (2024) find persona variables explain little annotation variance. Cui et al. (2025) replicate 156 scenario-based experiments with LLM subjects and find directional agreement with inflated effect sizes and high false-positive rates. Miranda and Balbi (2025) argue LLM survey predictions should be benchmarked against supervised models; we adopt that standard and find the strongest vanilla LLMs match or exceed a supervised demographics-only reference.

Persona-prompt design. Lutz et al. (2025) systematically compare sociodemographic prompting formats and find interview scaffolds beat flat instruction, and natural sentences beat key-value blocks. Our rendering sweeps (§**??**) replicate the direction of the interview-format effect but bound it: across sixteen renderer variants and a held-out validation split, *wording* moves individual fidelity within noise, while *content* (a single grounded ideology token) moves aggregate bias by a full point on the 7-point scale. Format matters at the margins; what is injected matters at the core.

Population-level simulation and alignment. A recent line of work targets the population level directly. Cao et al. (2025) fine-tune LLMs on first-token probabilities to minimize divergence between predicted and real response distributions on global cultural surveys—a *learning-based* route to the same goal our Study 2 reaches by *engineering*: auditing and calibrating the generative layer, with no LLM fine-tuning. The two are complementary, and we contrast them in §7. Zhao et al. (2025) formalize partial- and full-attribute simulation (PAS/FAS) and benchmark them across eleven datasets in the LLM-S^3 suite, positioning synthetic respondents as diagnostic rather than replacement instruments—a framing we adopt; our dispersion metric and generative-layer audit are natural extensions of such benchmarks. Yu et al. (2024) improve LLM election simulation with a multi-step pipeline integrating demographic, temporal, and ideological reasoning; read against our Study 1, their gains and our losses bracket the design space—structured intermediate reasoning can help where naive single-shot injection of an inferred ideology label hurts, because the label is *obeyed* rather than reasoned over (§4.4). Taday Morocho et al. (2026) evaluate persona conditioning on World Values Survey microdata at scale and find it frequently degrades alignment, with distortions concentrated in a minority of items and underrepresented subgroups; this independently corroborates our individual-level negative result and motivates cell-level rather than pooled evaluation. Finally, SPIRIT (Li and Conrad, 2026) recovers self-reported opinions from personas inferred from individuals' own social-media traces, outperforming demographic personas—consistent with our division-of-labor conclusion: individual-level gains require individually-measured grounding (cf. Park et al., 2024), while population-sampled grounding pays at the population level.

D Pilot results in full

Pilot run, N=200 stratified ANES voters, `gpt-4o-mini`, both arms, 100% parse coverage, deterministic seeds, every prompt and completion logged. Survey-weighted. Marginal floor (always predict the population mode/mean): vote accuracy 0.565, party-ID accuracy 0.060; both arms clear it, so the baseline reproduces the qualitative Argyle property.

Metric	Backstory baseline	Grounded persona
Vote accuracy (individual)	**0.680**	0.571
Party-ID accuracy (7-pt exact)	**0.289**	0.157
Party-ID MAE (0–6)	**1.91**	2.21
Party-ID individual corr.	**0.323**	0.144
Vote subgroup corr.	**0.716**	0.360
Party subgroup corr.	**0.805**	0.109
Vote subgroup MAE	0.210	**0.159**
Party subgroup MAE	0.966	**0.505**
|Republican share bias|	0.208	**0.098**
|Mean party-ID bias|	0.967	**0.029**

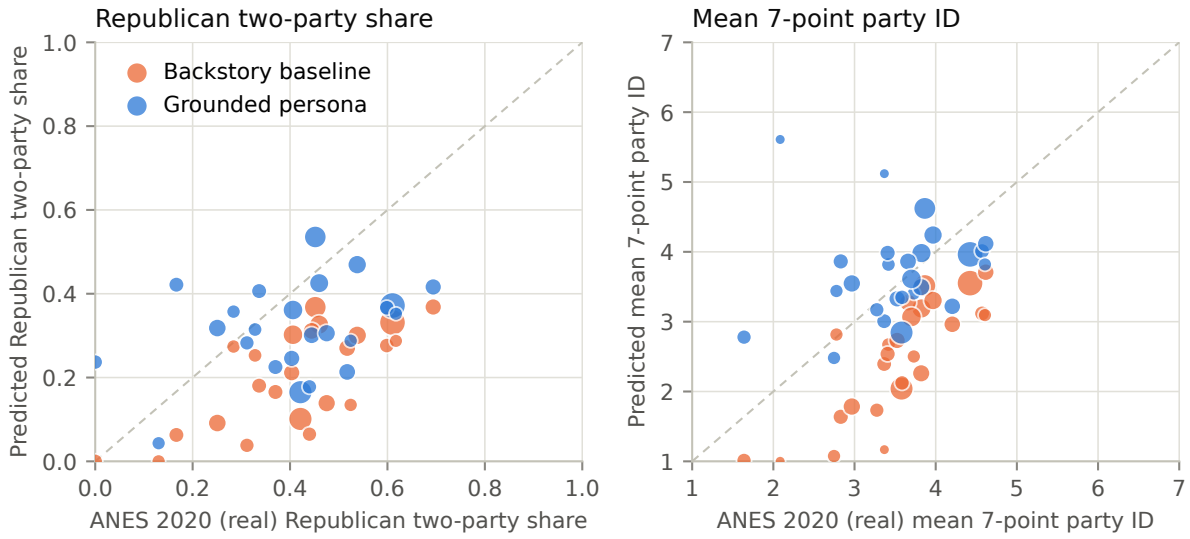

Figure 5: Pilot subgroup calibration (N=200, `gpt-4o-mini`). Each point is a demographic cell (sized by survey weight); the diagonal is perfect calibration. **Left:** Republican two-party vote share. **Right:** mean 7-point party ID. The backstory baseline (orange) is uniformly too Democratic but preserves ordering; the grounded arm (blue) is centered on the diagonal but noisier.

E Renderer variants (Study 1 sweeps)

Round 1: `v0` first-person mixed control; `v1` terse data sheet; `v2` rich narrative; `v3` numeric trait scores with anchors; `v4` demographics only; `v5` psychographics without ideology; `v6` ideology-forward; `v7` third-person instructional; `v8` survey framing. Round 2 (literature-informed): `v9` interview transcript; `v10` deep first-person narrative; `v11` verbal-trait variant of `v3`; `v12` structured sheet with numeric scores; `v13` interview + narrative combination. Validation (held-out N=120, three repeats pooled): `v9` was the only variant to beat control on both balanced accuracy (+0.014) and party correlation (+0.016); tuning-split calibration wins of `v10`/`v12` did not replicate. **Component ablation** (N=150, pre-repair artifact; objective = balanced accuracy + party-ID correlation penalized by bias): demographics only 0.670 bal. acc. / 0.36 bias on deepseek and 0.552 / 1.56 on gpt-4o-mini; adding ideology and psychographics gives 0.587–0.593 / 0.33–0.42 and 0.593–0.599 / 0.46–0.47; psychographics with ideology deleted is worst on both (0.494 / 0.90 and 0.510 / 1.64). Bare demographics win on the well-calibrated model; the grounded ideology field rescues the miscalibrated one.

F Fusion model specification

This appendix specifies the generative layer precisely enough to reimplement it. The model is deliberately simple: a weighted low-rank factorization with additive biases, chosen for inspectability and debuggability over expressiveness (its v2 successors—masked tabular transformers, conditional diffusion—are benchmarked against the same population contract).

Data structure. Training operates on a sparse observation matrix over rows i (a row is either an ACS backbone cell or a single survey respondent from a mounted source), fields f

Table 7: Cross-model comparison at the grounded arm's strongest configuration (repaired artifact, mode decode, label signal 0.251; elicitation temperature 0; N=150, N=60 for opus). Balanced accuracy and |mean party-ID bias|. Only sonnet's loss and gpt-4o-mini's win exceed the precision bound.

	Backstory baseline		Grounded (mode decode)	
Base model	bal. acc.	|bias|	bal. acc.	|bias|
claude-opus-4.8*	0.764	0.325	0.708	0.647
claude-sonnet-4.6	**0.688**	0.156	0.574	0.270
deepseek-chat-v3.1	0.659	0.182	0.615	0.506
llama-3.3-70b	0.604	0.381	0.583	0.351
gemini-2.5-flash	0.597	0.350	0.614	0.552
claude-haiku-4.5	0.573	0.524	0.572	0.532
gpt-4o-mini	0.530	1.324	**0.641**	**0.335**

(registered attributes: demographics, party, ideology, Big Five, moral foundations, attendance, trust, profession bridge fields), and sources s (ACS, CES, GSS, CPS, IPIP-FFM, AutoMoto, MFQ-2, Pol.is, Voteview, EJScreen). The production fit uses 4.87M observed entries; the runtime backbone exposes 270,479 ACS cells. Each observed entry carries an encoded scalar target y and a survey weight. Encodings: ordinal fields map to their level index $0, \ldots, K_f - 1$; bounded scores to $[0, 1]$; continuous fields are standardized with declared physical bounds clamped at decode; booleans to $\{0, 1\}$. (Unordered categoricals were historically encoded as an alphabetical index—the defect dissected in §5.2; politically consequential fields are now declared ordinal.)

Model. Parameters are row latents $z_i \in \mathbb{R}^d$ (here $d = 8$), field loadings $w_f \in \mathbb{R}^d$, field biases b_f, and per-source-per-field offsets $o_{s,f}$ that absorb instrument effects. The prediction for an observed entry is

$$\hat{y}_{i,f,s} \;=\; b_f + z_i^\top w_f + o_{s,f},$$

fit by weighted squared error. Three weighting mechanisms compose: (i) each entry's survey weight times a per-source reliability factor; (ii) *source/field balancing*—weights are renormalized so each (source, field) group contributes a fixed total, divided by $(\text{scale}_f)^2$ with $\text{scale}_f = K_f - 1$ for discrete fields, which stops large surveys and wide scales from dominating; (iii) *overlap anchors*: for every non-ACS row, its demographic overlap values are added as pseudo-observations with weight proportional to an anchor coefficient (0.2 in the production fit), tying survey-row latents into the ACS demographic field space—this is what makes latents comparable across sources.

Optimization. Stochastic gradient descent over shuffled entries; for entry e with error $\epsilon = \hat{y} - y$ and per-example step $\eta_e = \eta \cdot \min(\omega_e/\bar{\omega},\, 5)$:

$$z_i \mathrel{-}= \eta_e(\epsilon\, w_f + \lambda z_i), \quad w_f \mathrel{-}= \eta_e(\epsilon\, z_i + \lambda w_f), \quad b_f \mathrel{-}= \eta_e\, \epsilon, \quad o_{s,f} \mathrel{-}= \eta_e\, \epsilon,$$

with ℓ_2 coefficient $\lambda = 10^{-3}$, learning rate $\eta = 0.01$, up to 100 epochs, early stopping on held-out loss (the production quality profile disables the random split and gates on a full validation suite instead: source marginals, demographic slices, known within-source correlations, held-out lift). A batched PyTorch backend implements the same objective. Seeds fix initialization ($\mathcal{N}(0, 0.05^2)$) and shuffling.

Identifiability and missingness. The latent is identified only up to invertible linear reparameterization ($z \to Az$, $w \to A^{-\top}w$); no orthogonality or ordering constraints are imposed, and none are needed because only inner products enter predictions—the latent is never interpreted dimension-wise. Missing entries are simply absent from the objective (no imputation during training); at generation time every registered field is decoded from the latent regardless of source coverage, which is exactly the posterior-imputation behavior the validation suite flags per field as weak or strong support.

Generation. Given evidence (an audience specification), sampling proceeds:

```
1  compile evidence into criteria; split ACS-native vs. non-ACS
2  for each backbone cell c (prior weight π_c):
3    ACS-native criterion  → likelihood ∈ {0,1} (indicator on cell values)
4    non-ACS criterion     → smoothed support likelihood
     (matched weight + s)/(total weight + 2s) over survey rows
     overlap-compatible with c; model-based fallback if no support
5    posterior weight ∝ π_c ∏ likelihoods
6    cell latent z_c += λ_step (t_f − ŷ_f(z_c)) w_f/‖w_f‖  per non-ACS criterion
7  per persona: draw cell ~ posterior; z* = z_c + ε,  ε_d ~ N(0, (t·s_d)^2)
     (t = latent temperature; s_d = SD of row latents in dim d)
8  decode every field: x_f = b_f + ō_f + z*^⊤ w_f  (+ N(0, σ_f^2))
     ordinal w/ cutpoints → level #{j : c_j < x_f}; else round-and-clamp
9  clamp evidence fields to satisfy their criteria; run coherence passes
   (profession/life-stage consistency with the ACS cell)
```

where $\bar{o}_f$ is the field's mean offset over its observed sources, and all randomness is driven by deterministic per-sample seeds.

Uncertainty propagation (honest status). The decode is a point prediction plus the calibrated idiosyncratic noise of Appendix G; the per-field uncertainty surfaced to consumers is currently a two-level heuristic (lower for evidence-constrained fields), *not* a posterior variance. Propagating latent posterior uncertainty through the decode is future work; the dispersion-matching calibration is an empirical stand-in that restores population-level heterogeneity without per-draw uncertainty semantics.

G Decoder calibration: procedures

Both procedures operate on one ordinal field f with levels $0, \ldots, K-1$ and leave training untouched; their outputs (cutpoints $c_1 \le \cdots \le c_{K-1}$ and a noise scale σ_f) are persisted in the field's domain metadata and honored by every decode path.

The population prediction distribution. At operating latent temperature t, the sampler's raw prediction for field f on a persona from cell c is

$$X \;=\; \underbrace{b_f + \bar{o}_f + z_c^{\top} w_f}_{\text{cell base}} \;+\; \underbrace{\varepsilon_{\text{lat}}}_{\mathcal{N}(0,\ \sigma^2_{\text{proj}})} \;+\; \underbrace{\varepsilon_f}_{\mathcal{N}(0,\ \sigma_f^2)}, \qquad \sigma_{\text{proj}} = t\left(\textstyle\sum_d (s_d\, w_{f,d})^2\right)^{1/2},$$

since the per-dimension latent jitter projects through the loading. The population law of X is the π_c-weighted mixture over all backbone cells, estimated by Monte Carlo (50 draws per cell).

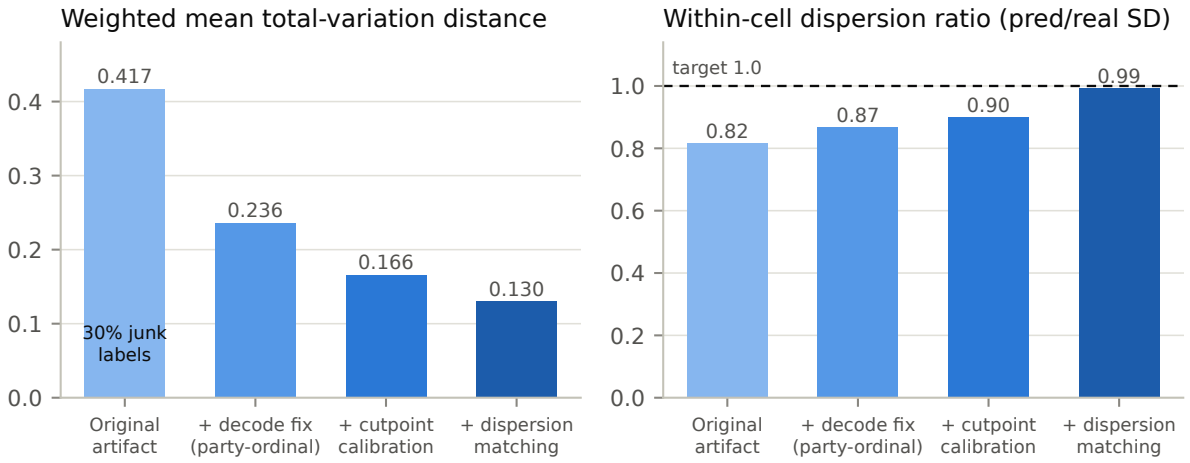

Figure 6: Fusion-only (no LLM) population fidelity across the repair sequence: 74 preregistered ANES cells, K=200 draws per cell, latent temperature 1.0. **Left:** survey-weighted mean total-variation distance to the real conditional party distribution. **Right:** within-cell dispersion ratio (predicted/real SD; 1.0 is perfect). The original artifact also emitted 30% junk labels, eliminated by the decode fix.

Cutpoint calibration (marginal matching). Given target shares $p_0, \ldots, p_{K-1}$ (the survey-weighted training marginal, CES+GSS pooled), set each cutpoint at the corresponding weighted quantile of the population prediction distribution:

$$c_j \;=\; \hat{F}_X^{-1}\Big(\textstyle\sum_{l \le j} p_l\Big), \qquad j = 1, \ldots, K-1,$$

with $\hat{F}_X$ the weighted empirical CDF. Decoding by $k(x) = \#\{j : c_j < x\}$ then reproduces the target marginal over the population by construction, up to Monte Carlo and atom granularity. Measured post-fit: party 44/39/18 vs. target 45/37/19; ideology within one point at every level.

Dispersion matching (conditional-variance restoration). The cell base is nearly deterministic given the cell, so without ε_f the within-cell variance of the decoded index is far below reality. The target is the true within-cell dispersion in the training population:

$$\tau_f \;=\; \frac{\sum_g m_g \,\mathrm{sd}_g}{\sum_g m_g},$$

over coarse cells g (the same eight one- and two-way specifications as the Stage-0 instrument, survey-weighted, mass ≥ 30), where sd_g is the weighted SD of the ordinal index within g. The fit bisects $\sigma_f \in [0, 3]$ (12 iterations) so that the *simulated* within-cell SD of the decoded index matches τ_f, refitting the cutpoints at every candidate σ_f (they depend on the noise), so the returned $(\sigma_f, c_{1:K-1})$ pair is self-consistent:

```
fit(f, shares, t, τ_f):
  measure(σ): X ← population predictions(t, σ); c ← quantile cutpoints(X,
shares);
          return within-cell SD of decoded index under (c, σ), c
  bisect σ ∈ [0,3] until within-cell SD = τ_f; return (σ, c)
```

Table 8: Joint-dependence audit: weighted Pearson correlation between decoded party (0–2) and ideology (0–4), persona-level (pooled draws over the 74 cells, K=200) and cell-level (across cell means). References: ANES 2020 voters (ideology from the 7-point self-placement) and the CES+GSS training population.

Configuration	Persona-level r	Cell-level r
ANES 2020 (truth)	0.777	0.907
CES+GSS training population	0.621	—
Decode fix, no calibration	0.319	0.732
+ cutpoints	0.337	0.743
+ cutpoints + independent dispersion noise	0.096	0.626
+ correlated dispersion noise ($\rho = 0.862$)	**0.607**	**0.751**

Fitted values: party $\tau = 0.882$ on the 0–2 axis $\rightarrow \sigma = 2.62$; ideology $\tau = 1.190 \rightarrow \sigma$ capped at 3.0 (slightly under-dispersed). Note the failure of the obvious alternative: estimating σ_f from training residuals gives 0.10–0.13, because per-row latents absorb conditional variance during training—residuals measure reconstruction error, not conditional dispersion.

What calibration does and does not touch—measured. Cross-field dependence flows exclusively through the shared cell identity and latent z^*; calibration re-maps each field's decode axis and adds *mutually independent* per-field noise, so pairwise associations between decoded fields are attenuated. We quantify this for the pair that matters politically (Table 8): persona-level party×ideology correlation is 0.78 in ANES 2020 voters and 0.62 in the CES+GSS training population; the uncalibrated decode carries 0.32, cutpoint calibration is harmless (0.34), and the dispersion noise collapses it to **0.10**. Cell-level structure survives better (0.63 vs. a true 0.91), which is why the cell-level results of §5.4 hold regardless. The measured verdict on independent noise: it buys marginal fidelity and honest within-field variance at a real cost in per-persona cross-field coherence—personas pair party and ideology implausibly more often than real respondents do. The remedy is a one-parameter extension, implemented and validated: draw the two fields' noise *correlated*, mixing the secondary field's standard normal as $\rho z_a + \sqrt{1-\rho^2}\, z_b$, which provably preserves each field's marginal and within-field dispersion (so the fitted cutpoints and σ remain valid unchanged). Fitting ρ by the same bisection pattern as σ—matching the sampler's within-coarse-cell decoded correlation to the training population's (0.615)—gives $\rho = 0.862$; end-to-end, persona-level coherence recovers to 0.607 (Table 8, last row), essentially the training-population reference, with cell-level structure improved and population TV/dispersion unchanged (0.133/0.986). The residual gap to ANES 2020 voters (0.777) is the training data's own weaker sorting plus the latent's limited coupling, which fusion-v2's coupling work (§5.6) addresses. **Stability:** cutpoints are a function of the operating point (t, σ_f) and are stored with it; changing the temperature requires a refit (deterministic given the seed). Bootstrap stability across matrix resamples has not been characterized.

H Same-data statistical baselines

Training population: 750,022 CES + GSS political rows carrying a party identification and the six cell demographics, survey-weighted by `person_weight`, party mapped to the fusion vocabulary by the same transform the matrix build uses. Scoring is the Stage-0 instrument: survey-weighted TV against the ANES-voter party conditional per cell (pid7 collapsed 1–3 / 4 / 5–7 to match

the training scheme), plus a within-cell dispersion ratio computed from K=200 draws from each predicted distribution so the metric is comparable to the sampled arms.

M0 is the pooled weighted marginal, returned unchanged for every cell. **M1** filters the training rows to the cell criteria and returns the weighted empirical party distribution; when a cell's matched training mass falls below 25, it drops constraints in a fixed order (income, education, age, region, race, sex) until the floor is met, recording what was dropped. **M2** is a weighted multinomial logit on one-hot demographics (30 features), fit by SGD with a decaying step and $\ell_2 = 10^{-5}$ over 25 epochs, and—critically—*marginalized* at prediction time: a cell is a subpopulation, not an individual at reference levels, so the fitted model is evaluated on every training row matching the cell and the predictions are weight-averaged. Scoring M2 at the cell's specified features alone, without marginalizing, produces a badly miscalibrated baseline (TV 0.31) and would have been a strawman.

I Run-log audit: a silently failed experiment

The first execution of the Stage-1 design (2026-07-16) failed silently. The API account's prepaid balance exhausted approximately 1,500 calls into the 29,896-call run; the harness's retry wrapper recorded each failed call with a sentinel value, the scorer excluded sentinels from aggregation, and the survey-weighted means were computed over whichever cells had data—4 of 74. The run printed per-arm summaries and reported completion; nothing in its output distinguished "all cells scored" from "4 cells scored." The resulting numbers were published in an earlier draft of this paper and evaluated by two external review passes before the defect surfaced.

It surfaced the same way the decode defects of §5.2 did: a per-cell integration of the verbalized-sampling arm compared cells one by one against the Stage-1 records, and 70 cells had nothing to compare against. The failure had been invisible at every aggregated view and was obvious at the first disaggregated one.

Three harness changes followed: (i) the account balance is preflighted against the estimated run cost before any call is queued (the mitigation the steering-stage harness already had—added after a previous balance exhaustion—but which the Stage-1 harness lacked); (ii) run summaries carry `n_failed`, per-arm `cells_with_data`, and an explicit `complete` flag; (iii) any failed call makes the run print an INVALID warning and exit non-zero, so a partial run cannot be scored by accident. The clean rerun of the identical design completed with zero failed calls; its full-population results are close to the accidental 4-cell ones, and the preregistered verdict was unchanged.

We report the episode at this level of detail for one reason: it is this paper's central methodological claim applied to ourselves. End-to-end summaries—of generative models and of experiment harnesses alike—blur the layer where defects live. *Score coverage, not completion counts.*

J Sensitivity and uncertainty

All analyses here are free: fusion-side sweeps plus resampling of the logged Stage-1 elicitations.

Sensitivity to K and latent temperature. Figure 7 (left) sweeps the fusion-only arm over $K \in \{10, 25, 50, 100, 200\}$ and latent temperature $t \in \{0.5, 0.75, 1.0, 1.25\}$: TV falls smoothly in K (finite-sample noise dominates below $K \approx 50$), $t = 1.0$ is best or tied at every K (confirming the operating point; the cutpoints are fit at $t = 1.0$, so departures also carry miscalibration), and the dispersion ratio is stable at 0.91–0.97 throughout. Figure 7 (right) subsamples the logged LLM

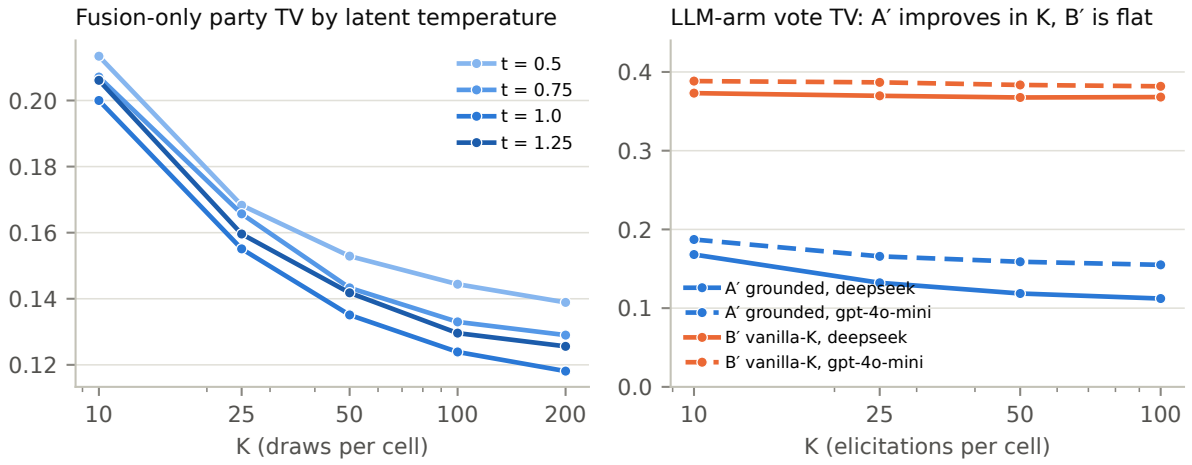

Figure 7: Sensitivity to K. **Left:** fusion-only party TV over the 74 cells by latent temperature; $t = 1.0$ (the operating point) is best or tied everywhere. **Right:** LLM-arm vote TV from subsampled logged elicitations: the grounded arm (blue) improves in K; vanilla-K (orange) is flat on both models.

elicitations: the grounded arm improves monotonically in K (deepseek vote TV 0.168 $\rightarrow$ 0.112 from K=10 to 100) while vanilla-K is *flat* (0.373 $\rightarrow$ 0.368)—a collapsed distribution cannot buy fidelity with more samples, which is the variance-collapse failure restated as a scaling law.

Survey-design uncertainty. The 74 cells are overlapping query families—six one-way marginals plus two crossings—so every respondent contributes to several cells, and the ANES conditionals are themselves survey estimates rather than fixed truths. Resampling cells would misstate both facts. We therefore compute intervals two ways, both of which recompute every cell conditional per draw and re-score the logged arm predictions against the recomputed truths: a respondent-level bootstrap, and a stratified cluster bootstrap on ANES's own variance design (resampling variance PSUs `V200010c` within strata `V200010d`), which respects the survey's clustering. The two agree closely; we report the design version. Vote TV: A′ 0.112 [0.105, 0.120] vs. B′ 0.368 [0.364, 0.372] on deepseek, and 0.155 [0.147, 0.165] vs. 0.382 [0.374, 0.391] on gpt-4o-mini; V 0.117 [0.114, 0.123] and 0.129 [0.126, 0.134]; D 0.041 [0.040, 0.049] and 0.045 [0.046, 0.053]. (The D-arm point sits at its interval's edge: truth resampling can only move a near-exact prediction's score up, a one-sided artifact of the design.) The H1′ separation holds under both resampling designs. One caveat these intervals do not carry: the arm predictions themselves are finite-K estimates and are held fixed here; Appendix J's K-curves bound that component separately.

Per-family results. No single aggregate carries the claim (deepseek vote TV, A′ / B′ / V / D): sex 0.124 / 0.484 / 0.097 / 0.058; race 0.188 / 0.401 / 0.122 / 0.048; age 0.064 / 0.329 / 0.108 / 0.028; education 0.075 / 0.422 / 0.109 / 0.037; income 0.089 / 0.427 / 0.111 / 0.018; region 0.084 / 0.304 / 0.148 / 0.032; education $\times$ race 0.169 / 0.280 / 0.098 / 0.043; age $\times$ region 0.104 / 0.295 / 0.140 / 0.063. The ordering is stable across all eight families. The grounded arm is weakest where the population is most heterogeneous within cell (race, education $\times$ race), which is also where verbalized sampling does best.

Pilot intervals. Respondent bootstrap for the Study-1 pilot (N=200): baseline balanced accuracy 0.674 [0.592, 0.755] vs. grounded 0.573 [0.476, 0.660]; baseline party-ID bias 0.967 [0.520, 1.394] vs. grounded 0.029 [0.006, 0.614]—the discrimination-versus-calibration split of §4.2 is interval-separated in both directions.

Precision of the cross-model study. Per-respondent logs were not retained for the seven-model comparison, so no bootstrap is available; as a worst-case bound, a proportion-style metric at N=150 carries a 95% half-width of at most ±0.08 (and ±0.13 at the one N=60 model). The directional pattern and the supervised-reference comparisons in Table 1 are robust to margins of that size; individual per-model differences smaller than ∼0.08 are not interpretable, and we do not interpret any.