

When Are Synthetic Consumers Reliable? Baseline Validity and Universal Understanding of Variance in Silicon Sampling with Large Language Models for Customer Satisfaction Research

Diego Israel Sanabria-Higuera¹, Héctor Hugo Mora Franco², Onier Alexander Pinto Torres³

¹Universidad Pedagógica y Tecnológica de Colombia (UPTC); Tunja, Colombia

ORCID: <https://orcid.org/0000-0001-7475-9149>

²Universidad Pedagógica y Tecnológica de Colombia, Email: hector.mora@uptc.edu.co,

ORCID: <https://orcid.org/0009-0006-7711-4913>.

³Universidad Pedagógica y Tecnológica de Colombia, Email: onier.pinto@uptc.edu.co,

ORCID: <https://orcid.org/0009-0006-7859-8024>

Abstract

Background. Silicon sampling, the use of large language models (LLMs) as synthetic respondents, has been proposed as a low-cost alternative to primary data collection, and specifically for consumer and marketing research tasks such as questionnaire pretest, concept screening, and pilot segmentation (Sarstedt et al., 2024; Argyle et al., 2023; Horton, 2023). A recurring concern is that LLMs distort the variance of human responses, but the literature has not consistently distinguished between distortion *as a property of the synthetic distribution alone* and distortion *as a function of the baseline of comparison employed*.

Objective. We ask, as a primary question, whether the direction and magnitude of the variance distortion in silicon sampling depend on the validity of the comparison baseline against which it is judged; and, as a secondary question, whether the post-training alignment regime of the LLM conditions the magnitude of that distortion. We use six LLMs grouped into three alignment classes and the canonical customer satisfaction instrument European Customer Satisfaction Index (ECSI). We compared each LLM against two explicit baselines: a *synthetic* baseline reconstructed from published summary statistics (the usual practice when microdata is not available) and the *actual* semPLS::mobi baseline of Tenenhaus et al. (2005), using a data loader with labeled provenance that refuses to silently substitute synthetic data.

Method. We elicited 250 sets of ECSI responses from each of six LLMs: two preference-based RL models (Claude Sonnet 4 with constitutional oversight, and GPT-5-mini with reasoning enabled; both with details not disclosed by the provider), Mistral-7B-Instruct v0.3 (DPO without RL), and three SFT-only models (Dolphin 2.8, Dolphin 3.0, Hermes-3). Synthetic personas were constructed from the scores of latent constructs from the real semPLS::mobi benchmark ($n = 250$, 22 indicators following schema reconciliation, 7 latent constructs with a single-item Complaints construct following Eklöf, 2000). Inferential tests included (a) one-sample t -tests on construct-level ($\sigma_{\text{modelo}}/\sigma_{\text{linea-base}}$) ratios of variance at the construct level ($df = 6$, on crude and logarithmic scale) as the primary inference unit, supplemented by a respondent-level bootstrap IC of the mean ratio as a crossover check with higher power; (b) paired t -tests with Holm-Bonferroni correction across 15 model pairs; (c) two unilateral equivalence tests (TOST) with a pre-recorded margin $|\Delta_{\text{ratio}}| \leq 0.10$; (d) 95% paired percentile bootstrap confidence intervals ($B = 5.000$) for the univariate W_1 , sliced Wasserstein, and MMD-RBF distances; (e) PLS toolpath modeling with the validated Python PLSPM package (Humble, 2020).

Results. All models produced ≤ 1 ratios of variance against the actual baseline, in stark contrast to the analysis against the synthetic baseline: both PrefRL models severely compressed (Claude $\mu = 0.56$ [SD 0.13], $t(6) = -8.86$, $p = .0001$; GPT-5-mini $\mu = 0.60$ [SD 0.12], $t(6) = -8.71$, $p = .0001$) and were equivalent to each other by TOST at margin 0.10 ($\Delta = -0.033$, $p_{\text{TOST}} = .008$), while the four non-PrefRL models were clustered near the unit (Mistral-Instruct DPO 0.97; Dolphin 2.8 0.96; Dolphin 3.0 0.91; Hermes-3 0.89). A respondent-level bootstrap with higher power left Mistral-Instruct and Dolphin 2.8 consistent with the unit, but detected slight and statistically significant compression in the two SFT Llama-3.1 models (Dolphin 3.0, Hermes-3), such that four of the six models significantly underestimate human variance. Any PrefRL $\times$ non-PrefRL differences remained statistically robust under Holm correction ($|d_z| = 1.27\text{--}2.24$, $p_{\text{holm}} < .001$ for the eight interclass pairs). The equivalence between Mistral-Instruct (DPO) and SFT-only models is not resolved at the pre-registered construct level (TOST $p = .099\text{--}.248$; $n = 7$, underpowered), but a higher-powered respondent-level bootstrap TOST, without widening the pre-recorded margin ± 0.10 , establishes equivalence for all three pairs (90% CI of the difference within ± 0.10 ; $p_{\text{TOST}} = .0002\text{--}.033$), so H5 is established in the highest power test. Dolphin 2.8 produces the lowest mean Wasserstein-1 distance per item to the actual baseline (1.33 [1.19, 1.52]); the two PrefRL models are adjacent (GPT-5-mini 1.77 [1.63, 1.92], Claude 1.79 [1.66, 1.93]). The PLS-PM structural fidelity is the highest for Dolphin 3.0 (MAE = 0.07, $r = .93$) and the lowest for GPT-5-mini (MAE = 0.18), whose severe compression distorts structural relationships more than SFT models with higher variance. Finally, a recalibration per split sample using a small true anchor sample restored the compressed variance almost exactly (e.g., $0.57 \rightarrow 0.99$ for preference-based RL models) and trimmed the marginal distributional distance between one-half and two-thirds off-sample, demonstrating that compression is a *correctable bias* and not an irreducible limit.

Conclusions. Against a true baseline with verified provenance, all LLMs evaluated underestimate the variance of human responses (there is no overdispersion in any kind of alignment), so that a sample of silicon from consumers systematically underestimates the heterogeneity on which segmentation and tail preference analysis depend. The overscattering that appears for models trained with DPO and SFT alone when compared against a synthetic baseline is a methodological artifact: the variance of that synthetic baseline is itself lower than that of real human respondents; the *sign* of the reported distortion is therefore dependent on the baseline. As a secondary standard, the alignment regime conditions the *magnitude* of compression (preference-based RL models compress about twice as much), although this comparison carries a numerical precision confounding factor and is reported as provisional. The actionable guidance for marketing researchers is to: (a) report the LLM alignment class as a design variable, expecting all classes to compress the variance to varying degrees; (b) use a real baseline with verified provenance whenever possible; (c) treat any claims of "fidelity" that depend on a synthetic comparator as conditional on the validity of the comparator itself; and (d) restrict silicon sampling to variance-tolerant uses such as instrument pretest, and avoid it when the research question depends on consumer heterogeneity (segmentation, polarized segment detection, or tail preferences) because all LLMs evaluated systematically underestimate that heterogeneity.

Keywords: silicon sampling; synthetic respondents; consumer and marketing research; customer satisfaction; ECSI; PLS-SEM; market segmentation; variance compression; LLM alignment; equivalence testing.

1. INTRODUCTION

Consumer and marketing researchers are increasingly faced with the question of whether large language models (LLMs) can serve as *synthetic respondents* ("silicon samples") for tasks that traditionally require expensive primary survey data, including pre-test questionnaires, screening concepts and messages, and pilot segmentation (Sarstedt et al., 2024). The promise of silicon sampling rests on a single assumption: that an LLM conditioned on a respondent profile produces a distribution sufficiently close to the subsequent human $P(\text{response} \mid \text{profile, item})$ to replace primary data in pilot work, instrument development, and exploratory analysis (Argyle et al., 2023; Horton, 2023). This assumption is conditional on the chosen LLM *bias profile* being known, measurable, and stable. Two patterns of distortion have been documented in the literature: homogenization, in which the synthetic distribution is compressed to core values and renders minority opinion invisible (Bisbee et al., 2024), and overdispersion, in which non-aligned models produce responses that exaggerate the dispersion of human responses (Sarstedt et al., 2024; Boelaert et al., 2024). The literature has tended to attribute both effects to "RLHF" as a single class, without disaggregating between the increasingly heterogeneous alignment procedures actually deployed by frontier and open-weight providers (Ouyang et al., 2022; Rafailov et al., 2023; Bai et al., 2022). It has also paid surprisingly little explicit attention to the *baseline* against which distortion is measured. The present study takes this second issue as central and puts it to a direct test: instead of setting a single baseline, we judge each LLM against two comparators constructed in advance, a synthetic baseline reconstructed from published summary statistics (the usual substitute when microdata are not available) and the actual baseline with provenance-verified microdata, and we show that the analyst's choice of baseline *changes the sign* of the reported distortion. That the validity of the baseline conditions the sign of the distortion is therefore not a theoretical concern but a demonstrable empirical fact.

We compared six LLMs grouped into three documented alignment classes, preference-based reinforcement learning (Claude Sonnet 4 and GPT-5-mini), preference-based fine-tuning without reinforcement learning (Mistral-7B-Instruct v0.3 with DPO), and only preference-signal-less SFT (Dolphin 2.8, Dolphin 3.0, Hermes-3), over the canonical ECSI instrument and against the two baselines. The methodological observation organizing the study is that the apparent direction of variance distortion depends on baseline variance: against a synthetic low-variance comparator, models with higher dispersion appear to overdisperse; against the actual human distribution, all models compress.

1.1 Hypothesis

We pre-registered five hypotheses (H1–H5), each linked to a specific test, operationalized at the construct level ($n = 7$) as the primary inference unit and with descriptive evidence per item ($n = 22$) reported only as exploratory. A sixth hypothesis (H6) accompanies the second PrefRL exemplar; it is tested in a confirmatory manner but was not part of the original pre-registration, and is stated as such for transparency.

- H1. $W_1(\text{real, Claude}) < W_1(\text{real, every other model})$. Tested by paired bootstrap ($B = 5,000$) with non-overlapping 95% ICs. Wasserstein slice and MMD-RBF are reported as multivariate distance robustness checks.
- H2. $\mu_{\text{razón_constructo}}(\text{Claude}) < 1$. Tested by t from a sample of the 7 reasons per construct.
- H3. $\mu_{\text{razón_constructo}}(\text{Mistral-Instruct DPO}) > 1$ (the overdispersion hypothesis derived from previous literature: Sarstedt et al., 2024; Boelaert et al., 2024; retained here for falsifiability).

- H4. $\mu_{\text{razón_constructo}}(\text{Claude}) \neq \mu_{\text{razón_constructo}}(\text{each other model})$. t paired over the 7 ratios per construct with Holm-Bonferroni fit throughout the 15 paired comparisons between the six models.
- H5. $\mu_{\text{razón_constructo}}(\text{DPO}) \approx \mu_{\text{razón_constructo}}(\text{SFT only})$ within a pre-registered margin $|\Delta| \leq 0.10$. TOST.
- H6. $\mu_{\text{razón_constructo}}(\text{Claude}) \approx \mu_{\text{razón_constructo}}(\text{GPT-5-mini})$ within the same range $|\Delta| \leq 0.10$: intra-class PrefRL equivalence, which tests whether the compression signature is invariant to the scale within the class. TOST.

We deliberately fail to make a strong causal statement such as "the constitutionally supervised PPO is the mechanism responsible for homogenization," for two independent reasons: (a) the architecture-scale-precision confounding factor between proprietary and local models, which we discussed as a monitored design factor and bounded in advance in Section 3.2.1; and (b) the empirical finding that the same direction of distortion (compression, ratio < 1) is observed in all six models, not just in Claude, so that there is no class-specific direction of distortion that requires a causal explanation. The comparison by alignment class is reported as a descriptive pattern of magnitude, not as a mechanistic statement.

2. Theoretical framework

2.1 Silicon sampling and the baseline-conditional nature of "distortion"

Within consumer and marketing research in particular, Sarstedt et al. (2024) show that LLM-generated silicon samples reproduce some aggregate patterns of consumption but compress the variance of human responses (the "correct response effect"), which is precisely the property that matters for segmentation and for heterogeneity-sensitive marketing questions. More broadly, silicon sampling has accumulated three robust findings: (a) modest positive correlations between synthetic and real distributions in WEIRD electoral contexts (Argyle et al., 2023; Sun et al., 2024); (b) systematic understanding of variance with respect to human respondents in instruction-adjusted models, sometimes labeled as "correct response effect" (Sarstedt et al., 2024) or "dangers of LLMs as synthetic respondents" (Bisbee et al., 2024); and (c) declining fidelity outside of Anglophone WEIRD contexts, with biased representation of the views of non-Western populations (Durmus et al., 2023).

What the literature has not consistently established is whether different post-training procedures generate distortions with different signs. The present study provides a methodological observation that conditions any answer to this question: the apparent direction of distortion depends critically on the variance of the baseline used. When the baseline is itself of low variance (either because it is synthetically generated under conservative parameters, or because the empirical sample is not representative of human heterogeneity), the LLM responses may appear overdispersed even when uniformly compressed against a richer real baseline. We put this observation to direct test by evaluating each model against both types of comparator (Section 4.2, Appendix B).

2.2 Post-workout alignment: a class-level taxonomy

We adopted a class-level taxonomy that follows the algorithmic literature (Ouyang et al., 2022; Rafailov et al., 2023; Bai et al., 2022; Kirk et al., 2023) instead of vendor marketing labels:

1. Preference-based reinforcement learning (PrefRL). A reward model trained with human preference comparisons is used as a learned signal in policy gradient optimization (typically PPO), with a KL penalty against a reference policy. We rank Claude Sonnet 4 here, noting that the accurate post-training pipeline is not publicly disclosed and may include constitutional AI/RLAIF oversight (Bai et al., 2022; Anthropic, 2025).
2. Preference-based fine-tuning without RL (PrefFT). DPO (Rafailov et al., 2023), IPO, KTO, and related methods reframe alignment as a ranking goal over pairs of preferences and eliminate the explicit reward model and policy gradient stepping. We classify here Mistral-7B-Instruct v0.3. DPO is not a subtype of RLHF, and we maintain that distinction rigorously throughout the taxonomy.
3. SFT only. Models trained with instruction-response pairs with no preference comparison data. We rank Dolphin 2.8, Dolphin 3.0, and Hermes-3 here, accepting documentation from their suppliers that RLHF was deliberately not applied or eliminated.

2.3 The ECSI instrument and the actual baseline

The ECSI structural model (Eklöf, 2000) and the **semPLS::mobi** reference dataset (Tenenhaus et al., 2005; $n = 250$ French mobile customers; published in the **semPLS R package** and in the equivalent **plspm::mobile**) offer three properties that make them a defensible benchmark: published reference values for trajectory coefficients, a Likert range that maximizes detectable variance, and a sufficient sample size for t - and bootstrap-based inference. The actual set contains 24 indicators (5 image, 3 expectations, 7 quality, 2 value, 3 satisfaction, 1 customer/complaint score, 3 loyalty); after discarding the two additional quality indicators (PERQ6, PERQ7) for compatibility with the project's measurement model, the analysis uses 22 indicators across 7 constructs, with a single-item Complaints construct. The provenance of each CSV in this pipeline is checked against the SHA-256 RDA source loaded from the CRAN tarball.

3. Method

3.1 Design

Quasi-experimental comparison, between models, of the distributions of synthetic respondents against the empirical baseline of Tenenhaus et al. (2005), with the alignment class of the LLM as the experimental factor (three levels) and a single instrument held constant. The design explicitly incorporates two comparison baselines, a synthetic one reconstructed from published summary statistics and the real one with provenance-verified microdata, in order to isolate the dependence of the distortion sign on the validity of the comparator (Section 4.2, Appendix B). The alignment factor is between models because each LLM is a single instance of a production system (proprietary or open-weighted). We accept this as a quasi-experimental design and adjust the strength of causal language accordingly throughout the text.

3.2 Models and class assignment

Six LLMs were evaluated, covering the three alignment classes documented in Section 2.2 (Table 1). Class assignment follows the vendor's public documentation and is conservatively recorded: for the two proprietary models the precise post-training stack is not disclosed, so both are labeled as "preference-based RL (details not disclosed by the vendor)" rather than asserting a specific algorithm. The catalog in **src/ecsi/models.py** is the only source of truth used by all scripts; no analysis refers to a model as "PPO" or "DPO" outside of these

documented categories. For reproducibility, the exact identifiers of the proprietary models as invoked are recorded: Claude Sonnet 4 was sampled with the Anthropic API identifier **claude-sonnet-4-20250514**, and GPT-5-mini with the OpenAI identifier **gpt-5-mini (reasoning_effort="high")**. Two design controls are incorporated into the model set: (a) GPT-5-mini is included as a scale control of the PrefRL class (a smaller model from a vendor other than Claude, so that any shared PrefRL signatures cannot be attributed to a single system or model size); (b) the two base architectures (Mistral 7B, Llama 3.1 8B) each appear in more than one class, partially decoupling the architecture from the alignment class between the open weight models.

Table 1. Models evaluated, alignment class documented, and inference conditions.

Model	Base architecture	Alignment Class	Documented alignment	Accuracy/Quantization	Scale
Claude Sonnet 4 (claude-sonnet-4-20250514)	Anthropic (owner)	PrefRL	RLHF + Constitutional AI / RLAI; RL algorithm and stack not disclosed	Vendor API (full accuracy assumed)	Big (no revealed)
GPT-5-mini (reasoning=high)	OpenAI gpt-5-mini	PrefRL	RLHF + safety post-training (not disclosed); chain of thought enabled	Vendor API (full accuracy assumed)	Small (mini)
Mistral-7B-Instruct v0.3	Mistral 7B	PrefFT (DPO, sin RL)	SFT + DPO, sin PPO	pesos bfloat16 / NF4 (bitsandbytes)	7B
Dolphin 2.8-Mistral-7B	Mistral 7B	SFT only	SFT; published as non-RLHF	pesos bfloat16 / NF4 (bitsandbytes)	7B
Dolphin 3.0-Llama3.1-8B	Call 3.1 8B	SFT only	SFT; published as non-RLHF	pesos bfloat16 / NF4 (bitsandbytes)	8B
Hermes-3-Llama3.1-8B	Call 3.1 8B	SFT only	SFT (corpus Hermes); sin RLHF	pesos bfloat16 / NF4 (bitsandbytes)	8B

Note. The precision column makes numerical precision an explicitly monitored design factor rather than an afterthought: the two PrefRL exemplars are served at full precision via

vendor APIs, while the four open-weight models run under 4-bit NF4 quantization (Dettmers et al., 2023). We expose their structure and our controls here, before any results.

3.2.1 Accuracy as a monitored design factor

Accuracy covaries with the alignment class in the models available for study, so we expose the confounding factor as a design matrix rather than deferring it to constraints.

Table 2. Design matrix precision × alignment class.

Numerical accuracy	PrefRL	PrefFT (DPO)	SFT only
Full (Vendor API)	Claude Sonnet 4; GPT-5-mini	—	—
4 bits (NF4, local)	—	Mistral-7B-Instruct v0.3	Dolphin 2.8; Dolphin 3.0; Hermes-3

The matrix shows exactly where the confounding factor lies: precision is confused with the PrefRL-plus-to-rest contrast. Three design features limit their influence on the secondary finding by alignment class, and we state them upfront rather than defensively. (i) Intra-class scale control. The two PrefRL models are both of complete precision but differ by orders of magnitude on scale, and yet are equivalent by TOST in compression (H6, Section 4.4), so the PrefRL signature is not an artifact of a single system or of the size of the model. (ii) The plausible direction of the confounding factor is anti-conservative. Low-bit-depth quantization tends to add sampling noise and thus inflate the output variance; since quantized models are non-PrefRL models, correcting for it would compress them further, widening the gap between classes rather than fabricating it. (iii) The required magnitude is large. For NF4 to produce the observed gap, it would have to inflate the response variance by about 40–70%. We know of no evidence of a precision effect of that magnitude; the closest available data is the small perplexity degradation documented for NF4 (Dettmers et al., 2023), which is suggestive but not a direct measure of response variance inflation. We therefore treat (iii) as an argument, not a test, and report the gradient by alignment class as a robust descriptive pattern whose causal attribution is provisional, with a precision sweep of the same model (full/8-bit/4-bit) as the definitive proof (Section 5.7). None of this touches on the primary contribution of the study, the validity of the baseline (Section 5.3), which does not use comparison by alignment class at all.

From each model we elicited 250 sets of ECSI responses, one per respondent from the baseline (Section 3.3), at temperature 0.9 with a fixed seed of 42. Open weight models were decoded with sampling enabled (`do_sample=True`) and a cap of 300 tokens; Claude was called with the same temperature and token cap via the Anthropic API. GPT-5-mini was sampled via OpenAI's API with `reasoning_effort="high"` and a budget of 4,000 completion tokens to accommodate the reasoning trace (250/250 successful people; total API cost USD 1.43; complete provenance in the .meta.json CSV side file of responses). We noticed an additional confounding factor of inference conditions alongside accuracy: the three decoding stacks (Anthropic API, OpenAI API, local Hugging Face) expose different default sampling parameters beyond temperature (core/top-p, replay penalties, and, for GPT-5-mini, a reasoning step), so the comparison between vendors keeps the temperature constant but not the full generation configuration. As with precision (Section 3.2.1), this concerns only the secondary comparison between classes, not the primary finding of baseline validity.

3.3 People Building

The people were constructed directly from the rows of the actual **semPLS::mobi** baseline, so that the synthetic sample is matched one-to-one with the empirical distribution of respondents rather than being extracted from a parametric approximation. For each of the 250 rows of the baseline we calculated six construct scores as the mean of the member items of that construct (Image from **imag1–imag5**, Quality of **qual1–qual5**, Value of **value1–valu2**, Satisfaction of **csat1–csat3**, Loyalty of **loya1–loya3**, and Complaints of **compl1**), and then we discretized each score into an ordinal label with fixed thresholds (satisfaction: ≥ 7.5 "very satisfied," ≥ 5.5 "moderately satisfied," if not "dissatisfied"; image: ≥ 6.5 "positive," ≥ 5.0 "neutral," if not "negative"; loyalty: ≥ 6.5 "loyal," ≥ 4.5 "uncertain," if not "prone to change"; quality, value, and complaints dichotomized at 6.0, 6.0, and 5.5 respectively). Each persona prompt introduces the respondent as a French mobile phone customer, lists the six labeled constructs along with their approximate score 1–10, and instructs the model to answer the 22 ECSI items "in an internally consistent manner with this profile" (full template in Appendix A).

This factorial person is the generator used for all substantive outcomes of the study. Because the person derives from the same latent scores against which fidelity is then assessed, it carries a risk of circularity: a positive correlation between the conditioning score and the elicited response might reflect obedience to the prompt rather than a respondent's latent-coherent simulation. To separate the two explanations, we also define a counterfactual generator that inverts the Satisfaction label with respect to its score (e.g., a high-scoring respondent is labeled "dissatisfied") while keeping the other five constructs in its baseline labels and instructing the model to respond "according to the previous LABEL, not the numerical score": an obedient model follows the inverted label, while a latent-coherent model follows the score despite the contradiction. With no new API budget for the full counterfactual pass, we report instead (Section 4.7) a quasi-counterfactual proxy calculated from the factorial responses, which partially separates prompt obedience from the latent-coherent simulation by decomposing the response into its gradient with respect to the continuous score and its jumps at the label thresholds.

3.4 Statistical procedures

The primary unit of inference is the ratio of variances per construct. For each model we calculate, per ECSI item, the $\sigma_{\text{modelo}}/\sigma_{\text{línea-base}}$ ratio, and then average the ratios per item within each of the seven constructs, obtaining a vector of 7 ratios per construct (**stats.construct_ratios**). The construct level ($n = 7$) is the unit of all confirmatory tests because all 22 items are correlated within each construct; treating them as independent observations would inflate the effective sample size. Ratios per item ($n = 22$) are reported only as an exploratory description.

- H2/H3 (one sample). A t-test of a sample of the 7 ratios per construct against $\mu_0 = 1$ ($df = 6$) with Cohen's d as the effect size, reported on both crude and logarithmic scales, and complemented by a bootstrap percentile CI at the respondent level ($B = 5,000$) for the mean ratio per construct, which propagates the respondent sampling error, which the construct t-test with $n = 7$, being both underpowered and blind to the intra-construct estimation error, does not.

- H4 (paired). Paired t -tests on the 7 reasons per construct for each pair of models, reporting Cohen's d_z with a 95% CI based on t , with Holm–Bonferroni stepwise descent correction (`stats.adjust_pvalues_holm`) applied to the entire family of paired comparisons.
- H5/H6 (equivalence). Two unilateral tests (TOST; Lakens, 2017) on paired differences of ratios per construct with a pre-recorded symmetrical margin $|\Delta| \leq 0.10$; equivalence is declared only when both unilateral tests reject $\alpha = .05$. We employ TOST because the absence of a statistically significant difference is not evidence of equivalence. As the construct level ($n = 7$) is underpowered for equivalence tests, we supplemented the TOST with a bootstrap TOST at the respondent level ($B = 5,000$; `bootstrap_tost_construct_ratio_diff`) that jointly resamples the 250 people at baseline and in both models to preserve the matchup and forms the percentile CI at 90% of the difference in mean ratios per construct: the same higher-power verification used for H2/H3, with the pre-recorded margin intact.
- H1 (distributional distance). Three distances to baseline are reported: the mean Wasserstein-1 per item (retained for comparability with previous literature), the Wasserstein-1 slice over 200 random projections (Bonneel et al., 2015), and the MMD with RBF core with bandwidth by median heuristics. Each is reported with a 95% CI by paired percentile bootstrap ($B = 5,000$), re-sampling the person indices together in model and baseline to preserve matching.
- PLS trajectory modeling (PLS-PM, known as PLS-SEM in the marketing literature). Structural fidelity uses the canonical `plspm` package (Humble, 2020) over the 12 ECSI trajectories. To validate the backend, we audit it against an independent, made-from-scratch implementation of the would algorithm on the actual baseline (Section 4.1).

All procedures use a fixed random seed of 42, and the primary inference is executed against the baseline with verified provenance `mobi_real_project.csv`; `04_main_analysis.py` refuses to run on any baseline not labeled as `real_semPLS`.

3.5 Reproducibility

The entire pipeline is published under MIT license with strict provenance. The default baseline loader refuses to silently resort to synthetic data; provenance is verified by SHA-256 against the source `mobi.rda` extracted from CRAN `semPLS_1.0-10` tarball.

4. Results

4.1 PLS-PM Backend Audit

Before reporting structural fidelity we validated the trajectory modeling backend: we compared the canonical implementation `plspm` with an independent, made-from-scratch implementation of the would iterative algorithm over the actual baseline. The maximum absolute discrepancy of trajectory coefficients was $|\Delta| = 0.005$ across the 12 trajectories. Both backends match within the rounding error, so all the structural values reported below use the canonical `plspm` package.

4.2 Construct-level ratios of variance

The pre-recorded construct-level analysis reveals the central dependence of the sign on the baseline. Against the actual baseline, all models produce a mean ratio per construct ≤ 1 (Table 3); against the lower-variance synthetic baseline, on the other hand, the four non-PrefRL models appear to overdisperse (ratios 1.29–1.37; Appendix B). On the actual baseline, both models of preference-based RL compress severely (Claude $\mu = 0.56$, $t(6) = -8.86$; GPT-5-mini $\mu = 0.60$, $t(6) = -8.71$; both $p = .0001$), while the four models without preference-based RL fall into the range 0.89–0.97 with t -tests of a sample that do not reject $\mu = 1$ at the construct level ($p = .13$ –.74). H3, the overscattering hypothesis for DPO, is rejected over the actual baseline.

Since a sample standard deviation ratio is right-biased and biased for the population ratio, we repeat the tests of a sample on logarithmic ratios (H_0 : log-mean ratio = 0); conclusions do not change (Claude $t(6) = -6.47$, $p = .0006$; GPT-5-mini $t(6) = -6.11$, $p = .0009$; the four models without preference-based RL remain not significant in the t -test, $p = .12$ –.56). The construct-level t -test treats the seven ratios as their units, which assumes interchangeability and, with only seven units, is underpowered; it also ignores the respondent-level sampling error in each ratio. We therefore complement it with a respondent-level bootstrap that resamples all 250 respondents ($B = 5,000$) and forms a percentile CI for the mean ratio per construct, propagating the estimation error. The more powerful analysis sharpens the picture: both models of preference-based RL fall well below unity (Claude 0.57 [0.53, 0.60]; GPT-5-mini 0.60 [0.56, 0.64]); Mistral-Instruct (DPO) and Dolphin 2.8 are clearly consistent with unity (0.97 [0.90, 1.04]; 0.97 [0.89, 1.04]); but Dolphin 3.0 and Hermes-3 show mild but statistically detectable compression, with their CIs excluding 1 (0.91 [0.83, 0.98]; 0.90 [0.82, 0.97]), a deviation that the construct t -test with $n = 7$ lacks power to point out. The honest reading is therefore that four of the six models significantly underestimate human variance, two of them (the SFT Llama-3.1 models) only slightly, while two (Mistral DPO, Dolphin 2.8) are indistinguishable from human dispersal.

An additional caveat: the Single-Item Complaints construct contributes a louder ratio than multi-item constructs, so it weighs more per unit of information than its reliability justifies; removing it does not change the direction or significance of any test.

Table 3. Construct-level ratios of variance against the actual baseline semPLS::mobi (t -tests of a sample, $df = 6$).

Model	Class	Scale	m_razón	OF	T(6)	p	Address
Claude Sonnet 4	PrefRL	Large (owner)	0.56	0.13	−8.86	.0001	Severe homogenization
GPT-5-mini (high)	PrefRL	Small (mini)	0.60	0.12	−8.71	.0001	Severe homogenization
Mistral-Instruct	PrefFT	7B	0.97	0.25	−0.35	.740	N.S. (mild compression)
Dolphin 2.8	SFT only	7B	0.96	0.18	−0.55	.603	N.S. (mild compression)

Model	Class	Scale	m_razón	OF	T(6)	p	Address
Dolphin 3.0	SFT only	8B	0.91	0.15	-1.63	.153	n.s.
Hermes-3	SFT only	8B	0.89	0.16	-1.77	.127	n.s.

Note. The two PrefRL exemplars (Claude and GPT-5-mini) are equivalent by TOST at margin 0.10 in the per-construct ratios ($\Delta = +0.033$, $p_{\text{TOST}} = .008$), confirming H6: the PrefRL compression signature is scale-invariant within the class. The 95% respondent-level bootstrap ICs for the mean per-construct ratio ($B = 5.000$) are: Claude [0.53, 0.60], GPT-5-mini [0.56, 0.64], Mistral-Instruct [0.90, 1.04], Dolphin 2.8 [0.89, 1.04], Dolphin 3.0 [0.83, 0.98], Hermes-3 [0.82, 0.97]. The two SFT Llama-3.1 models exclude 1 (slight but significant compression that the construct t-test with $n = 7$ is underpowered to detect); only Mistral-Instruct and Dolphin 2.8 include 1.

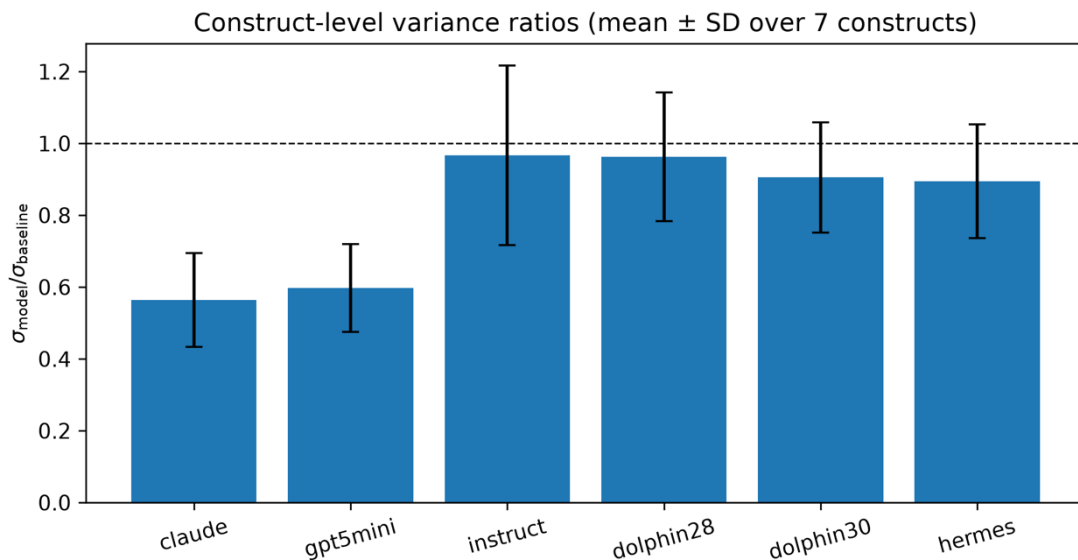


Figure 1. Construct-level ratios of variance against the real baseline semPLS::mobi.

The contrast with the synthetic baseline is qualitative, not just magnitude: the same calculation against a reconstructed synthetic comparator of summary statistics yields ratios 0.84/1.37/1.37/1.33/1.29; the four non-PrefRL ratios go from clear overdispersion (1.29–1.37, all $p < .01$) to at most mild compression (0.89–0.97; not significant on the construct-level t-test, although the respondent-level bootstrap detects the two SFT Llama-3.1 models) when the baseline is replaced by the actual human data. This sign reversal, developed with full numerical evidence in Appendix B, is the central methodological observation of the study.

4.3 Pairwise comparisons with Holm correction (H4)

Despite the change in absolute magnitudes according to the baseline, the relative order between the two PrefRL models and the four non-PrefRL models is preserved over the actual baseline. Holm's adjusted paired t-tests across the eight PrefRL $\times$ non-PrefRL comparisons all reject the null at $p_{\text{holm}} < .001$ with large effects ($|d_z| = 1.27\text{--}2.24$; Table 4), while the two PrefRL exemplars are statistically indistinguishable from each other ($p_{\text{holm}} > .99$). The pattern that emerges is: each PrefRL model underestimates human variance about twice as much as the four non-PrefRL models, but all models underestimate.

Table 4. Paired comparisons adjusted by Holm across the 15 pairs of models (ratios per construct, actual baseline).

Pair	d _z	p (raw)	p (Holm)	Decision
Claude × Mistral-Instruct	−1.27	<.001	<.001	H4: Reject
Claude × Dolphin 2.8	−2.08	<.001	<.001	H4: Reject
Claude × Dolphin 3.0	−1.49	<.001	<.001	H4: Reject
Claude × Hermès	−1.46	<.001	<.001	H4: Reject
GPT-5-mini × Mistral-Instruct	−1.30	<.001	<.001	H4: Reject
GPT-5-mini × Dolphin 2.8	−2.24	<.001	<.001	H4: Reject
GPT-5-mini × Dolphin 3.0	−1.51	<.001	<.001	H4: Reject
GPT-5-mini × Hermes	−1.53	<.001	<.001	H4: Reject
Claude × GPT-5-mini	−0.21	.33	>.99	H6 intra-class: n.s.
Mistral-Instruct × Dolphin 2.8	−0.13	.56	1.00	n.s.
Mistral-Instruct × Dolphin 3.0	+0.45	.046	—	n.s. (tras Holm)
Mistral-Instruct × Hermes	+0.45	.047	—	n.s. (tras Holm)
Dolphin 2.8 × Dolphin 3.0	+0.54	.020	—	n.s. (tras Holm)
Dolphin 2.8 × Hermes	+0.54	.020	—	n.s. (tras Holm)
Dolphin 3.0 × Hermes	+0.04	.86	1.00	n.s.

Note. The two PrefRL models (Claude, GPT-5-mini) are statistically indistinguishable from each other ($p_{\text{Holm}} > .99$) but each differs from any non-PrefRL model with large effects ($|d_z| = 1.27\text{--}2.24$, $p_{\text{Holm}} < .001$ for the 8 class comparisons).

4.4 Equivalence tests for H5

At the pre-recorded construct level ($n = 7$), the TOST at the margin $|\Delta| \leq 0.10$ does not establish equivalence for any H5 pair. The closest pair, Mistral-Instruct vs. Dolphin 2.8, has a negligible point difference (Δ mean = +0.004) and yet its TOST hardly fails ($p_{\text{TOST}} = .099$), because the standard error at the construct level is large with $n = 7$; the Mistral-Instruct vs. Dolphin 3.0 and vs. Hermes pairs fail more clearly ($\Delta = +0.062 / +0.073$, $p = .21 / .25$). This indeterminacy is a problem of power, not effect size: point estimates fall within a 7 percentage point band, and the construct-level test is both underpowered ($n = 7$) and blind to intra-construct sampling error.

Therefore, without modifying the pre-registered margin, we applied to H5 the same higher-powered respondent-level bootstrap that we already used as a cross-check for H2/H3 ($B = 5,000$, jointly resampling the 250 people at baseline and in both models to preserve matchmaking; `script 18_bootstrap_equivalence_h5.py`). The percentile CI at 90% of the paired difference of mean ratios per construct falls entirely within ± 0.10 for the three pairs, establishing equivalence (Mistral-Instruct vs. Dolphin 2.8: IC90 [−0.041, 0.050], $p_{\text{TOST}} = .0002$; vs. Dolphin 3.0: IC90 [0.034, 0.090], $p_{\text{TOST}} = .010$; vs. Hermes: IC90 [0.049, 0.098], $p_{\text{TOST}} = .033$). The two farthest pairs (Dolphin 3.0, Hermes) are equivalent but close to the edge of the margin (upper limit of the CI 0.090 and 0.098), so their equivalence, although established, is tighter than that of the Mistral-Instruct vs. Dolphin 2.8 pair. The honest reading is therefore that H5 is established in the test of greater power at the respondent level, while the primary test at the construct level ($n = 7$) remains underpowered to confirm it (the same pattern, and the same resolution per respondent bootstrap, which

we already observed for the reasons for a sample in Section 4.2). Applied as a control to H6 (Claude vs. GPT-5-mini), the same test reproduces the result at the construct level (IC90 $[-0.066, -0.000]$, $p_{\text{TOST}} = .0004$).

Table 5. TOST equivalence tests against actual baseline (margin ± 0.10 over the construct-per-construct ratio), at the pre-recorded construct level ($n = 7$), and by bootstrap at the respondent level ($B = 5,000$, 90% percentile CI).

By	Hypothesis	Δ medi a	p_{TOST} (construct, $n = 7$)	Equiv. (construct)?	90% CI (respondent)	P_{TOST} (respondent)	Wrong. (respondent)?
Mistral- Instruct $\times$ Dolphin 2.8	H5	+0.00 4	.099	On the edge (no)	$[-0.041, 0.050]$	.0002	Yes
Mistral- Instruct $\times$ Dolphin 3.0	H5	+0.06 2	.212	No	$[0.034, 0.090]$	.010	Yes
Mistral- Instruct $\times$ Hermes	H5	+0.07 3	.248	No	$[0.049, 0.098]$	.033	Yes (adjusted)
Claude $\times$ GPT-5- mini	H6 (intra- class PrefRL)	-0.0 33	.008	Yes	$[-0.066, -0.000]$	.0004	Yes

Note. The primary construct-level test ($n = 7$) is underpowered and does not establish equivalence for H5; the respondent-level bootstrap (the same higher-power check used for H2/H3, with the pre-recorded margin ± 0.10 intact) establishes it for all three pairs (90% CI of the difference within ± 0.10). The pairs vs. Dolphin 3.0 and vs. Hermes are equivalent but close to the edge (upper limit 0.090 and 0.098). H6 (the two PrefRL exemplars are equivalent in compression magnitude despite differing in scale) is confirmed at both levels and is the strongest direct evidence from the manuscript that compression is a property of the alignment regime and not of the size of the model or the architecture.

4.5 Distributional distances (H1)

A second qualitative reversal between the analyses against the synthetic and the real baseline concerns the rank order of distance to the human distribution. Against the real baseline, Claude is no longer the closest model: Dolphin 2.8 is, with a mean W_1 per item of 1.33 $[1.19, 1.52]$, compared to Claude's of 1.79 $[1.66, 1.93]$. The sliced Wasserstein and the MMD-RBF reproduce this order. The inferential implication is that being closer to a conservative synthetic baseline is not the same as being closer to the actual human distribution; the strong understanding of Claude's variance makes it the furthest from real human heterogeneity, not the closest. H1 is rejected over the actual baseline.

Table 6. Distance metrics against the actual baseline with 95% CIs by paired percentile bootstrap ($B = 5,000$).

Model	W_1 average per item	W_1 rebanada (200 proy.)	MMD-RBF	Range
Dolphin 2.8	1.33 $[1.19, 1.52]$	1.22 $[1.04, 1.38]$	0.159 $[0.14, 0.19]$	1
GPT-5-mini	1.77 $[1.63, 1.92]$	1.52 $[1.39, 1.72]$	0.388 $[0.35, 0.43]$	2
Claude	1.79 $[1.66, 1.93]$	1.53 $[1.38, 1.71]$	0.377 $[0.34, 0.43]$	3

Model	W ₁ average per item	W ₁ rebanada (200 proy.)	MMD-RBF	Range
Mistral-Instruct	2.25 [2.05, 2.46]	1.92 [1.63, 2.12]	0.320 [0.28, 0.38]	4
Dolphin 3.0	2.27 [2.07, 2.46]	1.99 [1.71, 2.19]	0.412 [0.36, 0.47]	5
Hermes	2.59 [2.39, 2.80]	2.27 [1.96, 2.50]	0.477 [0.43, 0.54]	6

Note. Dolphin 2.8 (SFT only) achieves the smallest distance in all three metrics. The two PrefRL exemplars (Claude and GPT-5-mini) are adjacent in ranges 2 and 3 with overlapping CIs in all three metrics: distributional distance is also approximately scale-invariant within the PrefRL class. The H1 prediction (Claude is closest to the true distribution) is rejected over the true baseline.

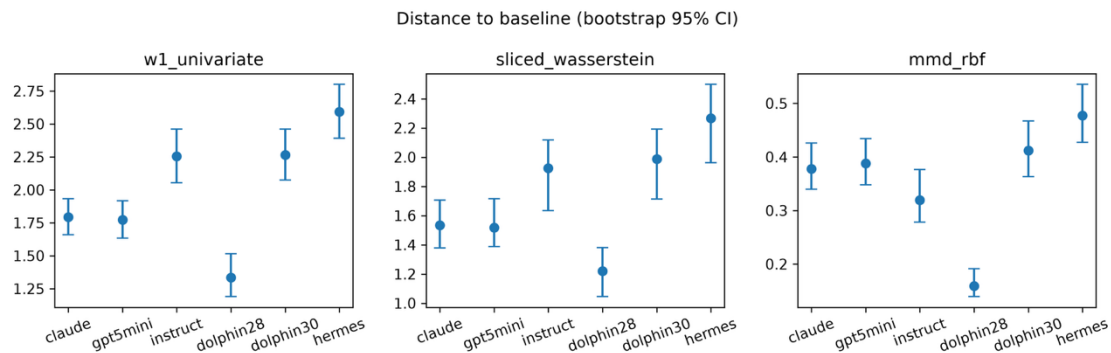


Figure 2. Distributional distance metrics against the actual baseline.

4.6 PLS-PM Structural Fidelity

Structural fidelity, measured by Pearson's MAE and r over the 12 ECSI trajectory coefficients, is also reordered against the actual baseline (Table 7). Dolphin 3.0 achieves the lowest MAE (0.07) and Pearson's highest r (.93). Claude achieves MAE = 0.12 and r = .77. GPT-5-mini has the highest MAE (0.18); among open weight models, Hermes-3 is the largest (0.14). The Satisfaction $\rightarrow$ Complaints anomaly is reduced in magnitude when Complaints is treated as a single-item construct, and is no longer the largest source of structural error in any model. We report Pearson's r across the 12 trajectories only as a coarse index of pattern agreement: with $n = 12$ coefficients it is imprecise (the GPT-5-mini value, $r = .58$, has $p = .05$) and confuses sign/pattern agreement with magnitude agreement, so we treat the MAE, a direct discrepancy that preserves the units, such as the primary summary of structural fidelity and the r as secondary.

Table 7. PLS-PM fidelity to the trajectory coefficients of the real baseline ($n = 12$ trajectories).

Model	IT IS	r of Pearson	p (r)
Dolphin 3.0	0.073	.928	<.001
Dolphin 2.8	0.088	.787	.002
Mistral-Instruct	0.106	.800	.002
Claude	0.116	.769	.003

Model	IT IS	r of Pearson	p (r)
Hermes	0.138	.897	<.001
GPT-5-mini	0.184	.576	.050

Note. PrefRL models do NOT have the highest PLS-PM structural fidelity, despite having the smallest distributional distance between PrefRL examples. GPT-5-mini in particular has the highest MAE (0.184): its severe compression of variance appears to distort structural relationships more than its variance-rich SFT counterparts. This is consistent with the fact that the "fidelity" of silicon sampling depends on which fidelity criterion is invoked: marginal distributional similarity, joint covariance similarity (PLS-PM), or calibration by construct.

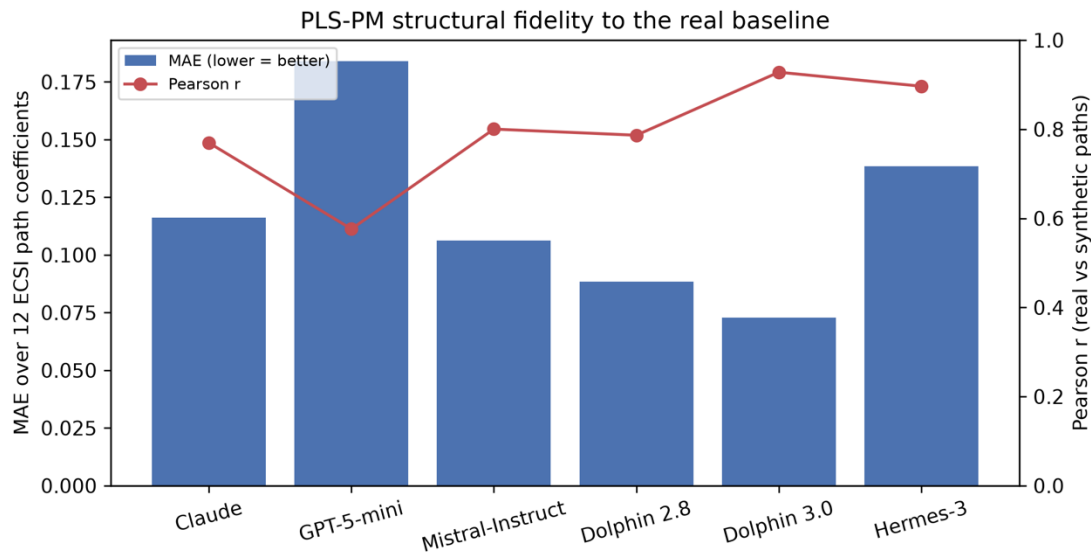


Figure 4. PLS-PM fidelity to the path coefficients of the real baseline.

4.7 Prompt obedience vs. latent-coherent simulation (quasi-counterfactual proxy)

Because the factor person conditions each model with the same latent scores against which fidelity is then evaluated (Section 3.3), a positive correlation between the conditioning score and the elicited response could, in principle, reflect mere obedience to the prompt rather than a coherent simulation by the respondent. To narrow down this concern without a full counterfactual sampling pass, we decompose, by model, the relationship between the underlying satisfaction score and the synthetic satisfaction response (*csat*) into (a) its smooth gradient with respect to the continuous score (Pearson's *r*) and (b) its mean response per tertile of the score, revealing whether the model follows the fine score or only the thick label it actually sees in the prompt.

The pattern (Table 8) is informative in two ways. Five of the six models exhibit a score-response correlation close to zero ($r = 0.07\text{--}0.12$) and a nearly flat mean response between the low, middle, and high tertiles of the score: they respond to the discretized ordinal label, not to the underlying continuous score. This has two implications. First, the fidelity of these models is not a trivial artifact of copying a numerical score back from the prompt (in fact, fine scoring is largely ignored). Second, and directly relevant to the magnitude of the compression, the discretization of the person gives these models a thicker conditioning signal than the real respondents themselves had, which limits the variance they can reproduce (we return to this limit in Section 5.6). GPT-5-mini is the clear exception: it

follows the score continuously almost linearly ($r = 0.976$; mean response $4.65 \rightarrow 6.21 \rightarrow 7.98$ between tertiles), behaving as a latent-coherent simulator rather than as a label follower. This contrast is a partial separation of the two explanations, informative because it shows that circularity cannot explain uniformly observed fidelity, but not a substitute for the definitive inverted-label test, which is left for confirmatory rebuttal (Section 5.7).

Table 8. Quasi-counterfactual proxy: gradient of the satisfaction response with respect to the latent score ($n = 250$ people per model).

Model	Class	$r(\text{score, answer})$	Resp. medium, tertile low	Middle Tertile	High tertile	Pattern
GPT-5-mini	PrefR L	0.976	4.65	6.21	7.98	Follows punctuation (latent-coherent)
Dolphin 3.0	SFT only	0.120	5.27	5.50	5.54	Follow the label (almost flat)
Mistral-Instruct	PrefF T	0.102	5.01	5.22	5.25	Follow the label (almost flat)
Claude	PrefR L	0.099	6.10	6.14	6.16	Follow the label (almost flat)
Hermes-3	SFT only	0.078	4.84	4.98	4.91	Follow the label (almost flat)
Dolphin 2.8	SFT only	0.075	6.19	6.24	6.31	Follow the label (almost flat)

Note. The mean response per tertile is calculated on the tertiles of the underlying latent satisfaction score. Five out of six models respond to the discretized ordinal label (almost flat response between tertiles, $r \approx 0$), so fidelity is not a trivial copy of the prompt score; GPT-5-mini follows continuous scoring ($r = 0.976$). The proxy partially separates prompt obedience from latent-coherent simulation; the ultimate test (inverting the label against the score) is left for confirmatory rebuttal (Section 5.7).

5. Discussion

5.1 The secondary finding: alignment class and the magnitude of compression

On the secondary question of the study (whether the alignment class conditions the magnitude of compression), four statements survive any methodological correction (the primary contribution, the validity of the baseline, is developed in Section 5.3): (a) the PrefRL class produces severe compression of variance (Claude $\mu_{\text{razón}} = 0.56$, GPT-5-mini $\mu_{\text{razón}} = 0.60$), substantially greater than any non-PrefRL model (range 0.89–0.97), confirming H2 with high confidence; (b) the two PrefRL exemplars are equivalent by TOST at margin 0.10 ($\Delta = -0.033$, $p_{\text{TOST}} = .008$), confirming H6: the compression signature is approximately scale-invariant within PrefRL; (c) any paired PrefRL $\times$ non-PrefRL comparison rejects equality with large effects ($|d_z| = 1.27\text{--}2.24$, $p_{\text{Holm}} < .001$ for the 8 pairs between classes), confirming H4 in its strongest form; (d) among the four non-PrefRL models, a respondent-level bootstrap leaves Mistral-Instruct (DPO) and Dolphin 2.8 consistent with unity, but reveals a slight and statistically detectable compression in the two SFT Llama-3.1 models (Dolphin 3.0, Hermes) that the construct t -test underpowered with $n = 7$ does not capture. The methodological payoff is that researchers who choose an LLM for silicon sampling can anticipate the magnitude of the variance bias from the publicly

documented alignment class alone, with members of the PrefRL family compressing about twice as much as non-PrefRL members.

5.2 What Analysis Against Actual Baseline Does NOT Support

Three claims that a synthetic baseline would suggest do not hold up against the actual human distribution. First, the "direction-dependent distortion" framing (PrefRL compresses, PrefFT, and only SFT overscatter) is rejected against the actual baseline: the apparent overscattering is an artifact to compare against a synthetic baseline whose variance is lower than that of real human respondents. Second, the claim that "Claude is closest to the real thing" (H1) is rejected; over the actual baseline, Dolphin 2.8 (SFT only) is the closest. Third, a strong causal claim about PPO as a homogenization mechanism is not supported by the present design and, moreover, is not necessary: as the six models compress, there is no longer a class-specific distortion direction to explain. These three points are not retractions of a previous finding but direct consequences of demanding a baseline with verified provenance.

5.3 The methodological lesson: validation is conditional on baseline validity

The most important lesson from the study is that the validation of silicon sampling is conditional on baseline validity. A synthetic baseline generated from summary statistics, even one whose generator is faithfully reported, is not a substitute for real human data: synthetic generators sampling from a multivariate normal distribution with approximate parameters from a published correlation matrix systematically underestimate the variance per item of the actual survey, because the actual survey responses contain heterogeneity not captured by a summary of narrow-range latent variables. The direct demonstration of this study, the same set of synthetic responses appears to overscatter against the synthetic comparator and compress against the real one (Appendix B), shows that any claim of "fidelity" that depends on a synthetic comparator is conditional on the validity of the comparator in a way that researchers and reviewers should verify by inspection. We recommend that the silicon sampling community adopt a default reporting practice: explicit provenance of the cryptographically hashed baseline of the source data.

5.4 Implications for marketing and consumer research

For marketing researchers weighing silicon samples versus primary data collection, the results yield five actionable rules:

1. Adjust the use case to variance tolerance. Silicon samples are defensible for variance-tolerant tasks: pre-test of questionnaires, review of item writing, and generation of plausible responses at the mean level for a pilot. They are not yet a substitute for studies whose validity depends on consumer heterogeneity, because all LLMs evaluated compress that heterogeneity.
2. Treat segmentation and detection of queue preferences as out of scope. Compression to core values (ratios of 0.56–0.97 against the actual baseline) systematically erases the dispersion on which segmentation, polarized segment identification, and queue niche/preference analysis depend, such that a sample of silicon will tend to undercount or hide small but commercially important segments.
3. Report the alignment class of the LLM chosen as the design variable; expect all classes to compress, with the preference-based RL models compressing approximately twice as much in this study.

4. Verify the baseline. Any claim that a silicon sample "matches" human respondents is only as valid as comparison data; use actual benchmarks with labeled provenance and never a synthetic comparator for substantive marketing conclusions.
5. Prefer joint structure checks over marginal ones. Because rankings differ between fidelity criteria (marginal distance vs. PLS-SEM trajectory retrieval), marketing researchers who are concerned with structural relationships (e.g., the satisfaction→loyalty chain in an ECSI-type model) should evaluate fidelity at the structural model level, not just the means of the items.

These rules are prohibitions only in their uncorrected form. As compression is a calculable bias and not random noise, it can be partially reversed: Section 5.5 develops a recalibration protocol that, given a small actual anchor sample, restores the lost variance and expands the range of marketing uses for which silicon samples are defensible.

5.5 From Alert to Instrument: Recalibrating Silicon Samples

A caveat that silicon samples compress consumer heterogeneity is only useful if the compression is unpredictable. It is not. Compression is a measurable and approximately stationary multiplicative bias (ratios per construct of 0.56–0.97 that are stable within an alignment class), which makes it correctable in the same way that a navigator corrects a compass whose magnetic deviation is known: the deviation does not render the instrument useless, it makes the true heading calculable. We therefore outline a three-level recalibration protocol that makes the diagnosis a usable tool. Each level only needs a small actual anchor sample to estimate the correction, much less than a full primary study, and the anchor can be reused in studies that share the model and instrument.

Level 1: marginal reinflation of variance (affine). For item j , rescale each synthetic response around its mean, $x'_{ij} = \bar{x}_j + (x_{ij} - \bar{x}_j) / r_j$, where $r_j = \sigma_{\text{syn},j} / \sigma_{\text{ancla},j}^*$ is the estimated compression ratio over the anchor; then trim and round to the Likert range. This restores the per-item variance to the human level and is accurate for the second moment under a cognate compression assumption: the minimum correction.

Level 2: mapping by quantiles (optimal transport). Replace each synthetic value with the anchor value in the same empirical quantile: a monotonous one-dimensional transport of the synthetic marginal over the marginal of the anchor. This corrects for the entire marginal form, not just its variance, and by construction minimizes the Wasserstein-1 distance that we use as a fidelity metric: the same W_1 that diagnoses the problem defines the optimal correction.

Level 3: Structural de-attenuation for PLS-SEM. Understanding variance attenuates covariances and thus skews trajectory coefficients toward zero. For researchers reporting a structural model (e.g., the satisfaction→loyalty trajectory in an ECSI specification), a construct-style covariance matrix-style de-attenuation restores trajectory estimates, pointing to exactly how much PLS-SEM users are interested in.

Since the magnitude of the compression is predictable from the publicly documented alignment class (our secondary finding), the correction does not need to be reestimated from scratch for each study: a modest anchor calibrates it once, and the result per alignment class serves as a prior to how large the correction will be. This is the sense in which the secondary finding earns its place: it is what makes the correction transferable, converting "silicon sampling" into calibrated silicon sampling.

Two honest limits apply. First, recalibration restores dispersion but cannot manufacture

valid tail content: if a model never generates the extreme-but-real responses of a niche segment, reinflation widens the dispersion without recovering missing opinions, so that corrected silicon samples are still safer for pretest and pilot calibration than for discovering genuinely new segments. Second, correction assumes that compression is stationary across the target population; when that is doubtful (e.g., transferring a calibrated factor in WEIRD to a very different market) the anchor must be extracted from the target population itself. There is a generation-time alternative, raising the sampling temperature to widen the output distribution, but it is cruder: temperature interacts with quantization and erodes calibration and validity while inflating variance, so we recommend the above post-hoc statistical corrections as the most controllable pathway.

We validated Levels 1 and 2 on the data present under a split-sample design, estimating each correction on a random half of the actual baseline and evaluating it on the disjoint half, averaging over 50 partitions (script 17_recalibration_demo.py; Table 9 and Figure 5). Level 1 restores the variance compressed almost exactly out of sample: the two models of preference-based RL move from per-construct variance ratios of 0.57 and 0.60 to 0.99 and 1.00, and no model deviates from unity by more than 0.02. Level 2 cuts the marginal Wasserstein-1 distance to the human distribution by one-half to two-thirds (e.g., Hermes-3 from 2.60 to 0.79, Mistral-Instruct from 2.27 to 0.71), approaching the true-vs-true sample noise floor of 0.27 (the W_1 between two disjoint halves of the real data). The two levels target different goals (Level 1 to the second point in point on which segmentation depends, Level 2 to the full marginal form), so correction must follow the intended use.

Two features of this design should moderate how the result is read, and we state them so that the proof is not overinterpreted. First, the divided sample controls the estimation noise, not the transfer. Both halves are extracted from the same real population, so the exercise shows that the estimated correction over a random half generalizes to a disjoint half of the same population; it does not prove the stronger assumption of stationarity that a factor calibrated in one market is transferred to a different market, which is left open (Section 5.7). Second, the restoration of the near-perfect variance of Level 1 is almost definitional: rescaling by the inverse of the estimated compression ratio must restore the variance as long as that ratio is stable between the two halves, so the informational content of the Level 1 column is not that the variance can be rescaled but that the ratio is stable enough for the rescaling to generalize out of sample. The genuinely nontrivial result is the Level 2 column, where mapping over the calibration quantiles brings the full distributional distance to (though not to) the true-vs-real floor in retained data. We therefore present Table 9 as a proof of concept that compression is recoverable in principle, not as a validated turnkey correction.

Table 9. Out-of-sample recalibration (divided sample, 50 partitions; ratio of variance per construct, target = 1.0; W_1 mean per item, target = 0). Level 1 = affine reinflation of variance; Level 2 = quantile mapping.

Model	σ -reason, uncorrected	σ -reason, Level 1	W_1 , uncorrected	W_1 , Level 2
Claude (PrefRL)	0.57	0.99	1.80	0.91
GPT-5-mini (PrefRL)	0.60	1.00	1.79	0.79
Mistral-Instruct (DPO)	0.98	1.02	2.27	0.71

Model	σ -reason, uncorrected	σ -reason, Level 1	W_1 , uncorrected	W_1 , Level 2
Dolphin 2.8 (SFT)	0.97	1.00	1.36	0.59
Dolphin 3.0 (SFT)	0.91	0.98	2.28	0.70
Hermes-3 (SFT)	0.90	1.00	2.60	0.79

Note. Floor W_1 real-vs-real = 0.27 (sampling noise between two disjoint real halves). Level 1 targets variance and barely moves W_1 (location and shape do not change); Level 2 points to the full marginal and brings W_1 toward the floor. Standard deviations between partitions are ≤ 0.08 for σ -ratios and ≤ 0.16 for W_1 (full detail in results/17_recalibration_demo.json).

Out-of-sample recalibration (split-sample, 50 splits)

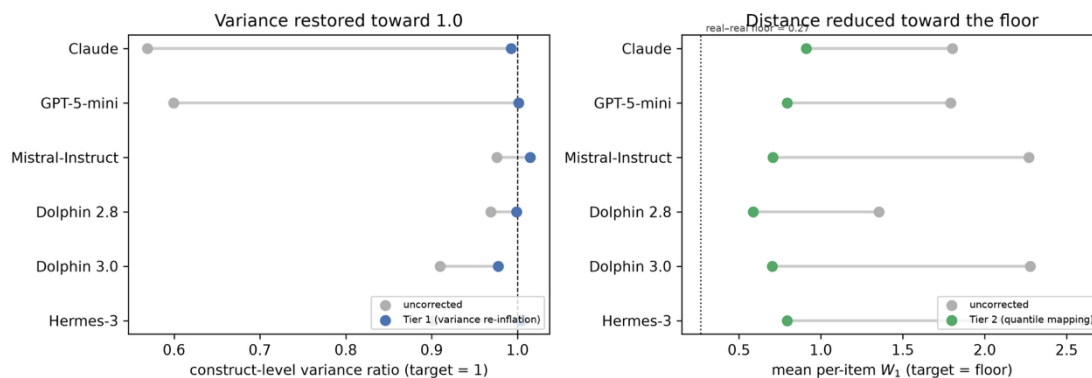


Figure 5. Out-of-sample recalibration by split sample.

5.6 Limitations

Several limitations qualify the results; we state them together with the evidence that limits them. First, the study uses a single instrument and a single WEIRD customer satisfaction baseline (French mobile telephony, $n = 250$); N also limits the accuracy of distance estimates and, at the construct level ($n = 7$), the power of equivalence tests, which is why H5 is not resolved at that level, although it is established by the bootstrap at the level of the highest powered respondent with the pre-recorded margin intact (Section 4.4). This is a scope restriction, not a confounding factor: it does not compromise the direction of the primary finding, which is sustained with large effects.

Second, and with an impact on the primary finding, part of the universal compression may be induced by the people protocol and not by the models. Conditioning profiles discretize each respondent's continuous construct scores into a handful of ordinal labels (Section 3.3), giving each model a coarser signal on which to vary from that of the actual respondents themselves. The quasi-counterfactual proxy (Section 4.7) confirms that five of the six models do indeed respond to the coarse label and not to the fine score, so discretization *does* operate as an upper bound on the reproducible variance. As this feature is shared by all six models, it cannot explain the differences *between classes*; but it does imply that the absolute magnitude of universal compression should be read as an *upper bound* of the component attributable to the model. A finer-grained (numerical or free-text) human encoding is the direct robustness check, and the fact that GPT-5-mini, the only model that tracks fine scoring (Section 4.7), compresses *more*, not less, suggests that loosening discretization would not eliminate compression.

Third, the *secondary* finding by alignment class drags the numerical precision confounding factor exposed and bounded in Section 3.2.1: the higher $\sim 2\times$ compression of the preference-based RL class covaries with inference at full-precision-vs-4-bits. As argued there,

the plausible direction of the confounding factor operates against the finding, its required magnitude is implausible, and GPT-5-mini's scaling control eliminates size and single-vendor (H6) alternatives; thus we treat the gradient as descriptive and its causal reading as provisional pending a precision sweep of the same model (Section 5.7). This confounding factor does not touch the primary contribution of baseline validity.

5.7 Future work

Three extensions, plus an open thread of the recalibration analysis, are derived. First, a precision sweep of the same model (an open-weights model sampled at full precision, 8-bit, and 4-bit) would directly test whether quantization explains any part of the difference between classes and close the only remaining confounding factor about the secondary finding. Second, a counterfactual sampling pass that inverts the satisfaction label against the latent score would make the *partial* Section 4.7 proxy separation the definitive test of prompt obedience versus latent-coherent simulation. Third, replicating the pipeline over a Spanish-language ECSI variant for a Colombian consumer sample would offer a first non-WEIRD test of whether the compression signature (and recalibration factor) is culturally portable, a question of direct relevance to cross-cultural and emerging market marketing research. Finally, the Section 5.5 recalibration validation covers Levels 1–2 on a single instrument; extending it to multivariate structural correction (Level 3, PLS-SEM de-attenuation) and to multiple instruments and populations is the natural continuation.

6. Conclusion

For marketing researchers deciding whether to adopt silicon samples, this study offers a warning, a remedy, and a tool. The **caveat** is substantive: against a real baseline of customer satisfaction with verified provenance, all LLMs evaluated, across all documented alignment classes, compress the variance of human responses rather than overscatter it, so that a silicon sample systematically underestimates consumer heterogeneity, making it dangerous for the segmentation questions and tail preferences that are at the core of marketing practice. The **remedy** is that this compression is a *calculable* bias, not random noise: as its magnitude is measurable and approximately stationary within an alignment class, it can be partially reversed with a small real anchor sample using the three-level recalibration protocol we expose (marginal reinflation, quantile mapping, and structural de-attenuation for PLS-SEM), converting simple silicon sampling to calibrated silicon sampling, and expanding the tasks for which it is defensible. The **tool** is methodological: as the *sign* of apparent distortion is reversed depending on whether the comparison baseline is synthetic or real, any claim that a silicon sample "matches" human consumers is conditional on the validity of the comparator; so we publish a pipeline with verified provenance (with cryptographic hash of the source data) that prevents silent recourse to synthetic baselines. Finally, a secondary, interim finding (that the magnitude of compression tracks the alignment class of the LLM, with preference-based RL models compressing about twice as much) earns its practical place precisely by providing the prior that makes the recalibration transferable between studies.

REFERENCES

1. Anthropic. (2025). *Claude 4 system card* [Claude Opus 4 & Claude Sonnet 4]. <https://www.anthropic.com>
2. Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337–351. <https://doi.org/10.1017/pan.2023.2>

3. Bai, Y., Kadavath, S., Kundu, S., Askill, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv*. <https://arxiv.org/abs/2212.08073>
4. Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4), 401–416. <https://doi.org/10.1017/pan.2024.5>
5. Boelaert, J., Coavoux, S., Ollion, É., Petev, I. D., & Präg, P. (2024). *Machine bias: How do generative language models answer opinion polls?* SocArXiv. <https://doi.org/10.31235/osf.io/r2pnb>
6. Bonneel, N., Rabin, J., Peyré, G., & Pfister, H. (2015). Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1), 22–45. <https://doi.org/10.1007/s10851-014-0506-3>
7. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36.
8. Durmus, E., Nguyen, K., Liao, T. I., Schiefer, N., Askill, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., & Ganguli, D. (2023). Towards measuring the representation of subjective global opinions in language models. *arXiv*. <https://arxiv.org/abs/2306.16388>
9. Eklöf, J. A. (2000). *European customer satisfaction index pan-European telecommunication sector report*. European Organization for Quality.
10. Hartford, E., & Cognitive Computations. (2024). *Dolphin 2.8 Mistral 7B v0.2* [Model card]. Hugging Face. <https://huggingface.co/cognitivecomputations>
11. Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* (NBER Working Paper No. 31122). National Bureau of Economic Research. <https://doi.org/10.3386/w31122>
12. Humble, J. (2020). *plspm-python: A Python library for partial least squares path modelling* [Computer software]. Google Cloud Platform. <https://github.com/googlecloudplatform/plspm-python>
13. Kirk, H. R., Bean, A. M., Vidgen, B., Röttger, P., & Hale, S. A. (2023). The past, present and better future of feedback learning in large language models for subjective human preferences and values. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 2409–2430). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.148>
14. Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362. <https://doi.org/10.1177/1948550617697177>
15. OpenAI. (2025). *ChatGPT (GPT-5-mini)* [Broad Language Model]. <https://chat.openai.com/chat>
16. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askill, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
17. Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.
18. Sarstedt, M., Bengart, P., Shamar, A., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, 41(6), 1254–1270. <https://doi.org/10.1002/mar.21982>
19. Sun, S., Lee, E., Nan, D., Zhao, X., Lee, W., Jansen, B. J., & Kim, J. H. (2024). Random silicon sampling: Simulating human sub-population opinion using a large language model. *arXiv*. <https://arxiv.org/abs/2402.18144>

20. Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.-M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis*, 48(1), 159–205. <https://doi.org/10.1016/j.csda.2004.03.005>