ECONOMIC-INFORMED PERSONA GENERATION FOR BEHAVIORALLY GROUNDED LLM POPULATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly used as persona-conditioned user simulators, yet a plausible persona prompt is not a behavioral specification: the same text can induce different choice patterns across simulators, and matching observed behavior may not identify responses to interventions. We introduce Economically-Informed Persona Generators (EIPG), a framework that calibrates persona generators from discrepancies between simulated and observed behavior rather than from persona labels. In a controlled benchmark, weak intervention coverage can improve calibration while worsening unseen counterfactuals; richer interventions make the calibration signal more identifying and improve frozen test performance. For natural-language personas, an anchored parameterization preserves authored prompts while exposing sparse, interpretable preference adjustments. Across four local LLMs from three model families, useful corrections are model-dependent, while structured economic search transfers more reliably than more flexible LLM- or text-evolution search in our finite-budget experiments. On public discrete-choice data, calibration improves an initially misspecified synthetic population on held-out human choices, although direct human estimation remains substantially stronger on the primary held-out choice metrics. EIPG therefore reframes persona grounding as an identification problem: the goal is not merely to write plausible personas, but to construct synthetic populations whose behavior remains credible under interventions.

1 INTRODUCTION

Large language models (LLMs) are increasingly used as persona-conditioned user simulators for recommender evaluation, market research, surveys, and synthetic preference generation (Park et al., 2023; Horton et al., 2023; Paglieri et al., 2026; Argyle et al., 2023; Aher et al., 2023; Brand et al., 2023). A typical deployment starts from descriptions such as budget-sensitive, quality-oriented, or sustainability-oriented and asks an LLM to act on behalf of each segment. The outputs may be fluent and plausible, yet plausibility does not guarantee that the induced synthetic population reproduces the behavioral tradeoffs needed for downstream analysis.

The core difficulty is simple: *persona text is not behavior*. The same prompt can induce different choice patterns in different simulators, and the same observed choices can often be explained by different underlying preference populations. Both problems matter for counterfactual use. A simulator may match baseline purchases while predicting the wrong substitution pattern when price, quality, or sustainability changes. In our experiments, for example, Gemma 3 4B and Gemma 3 12B receive the same quality-oriented prompt but calibration selects opposite corrections to brand sensitivity. The deployed simulator therefore mediates how persona language becomes behavior, while observed baseline fit alone need not identify the behavioral responses that matter under intervention (McFadden, 1980; Train, 2009).

We address this problem by treating the *persona generator itself* as a calibratable component of the simulation system. Economically-Informed Persona Generators (EIPG) close a feedback loop between persona construction, simulated behavior, and interpretable behavioral evidence. Rather than supervising a "correct" persona description, EIPG compares simulated and target quantities such as choice shares, chosen attributes, substitution patterns, and responses to controlled interventions. Structured generators can be calibrated through an inner economic model, while natural-language

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

personas can be calibrated directly from the behavioral moments they induce. In both cases, the object being grounded is the synthetic population's behavior, not the surface form of its description.

This framing makes *identification* the central technical issue. In a controlled benchmark, a weak price-only objective can improve during calibration while held-out counterfactual error becomes worse. Replacing a homogeneous multinomial-logit summary with a richer latent-class model does not repair the failure. What does help is more informative intervention coverage: perturbing price, quality, sustainability, and novelty makes the calibration objective more predictive of unseen behavior. After the richer specification is frozen, EIPG reduces final counterfactual error by 6.6% relative to the initial generator, versus 1.1% for the price-only comparator. The result is not simply that a more expressive model fits better; the evidence supplied to the calibration loop determines which behavioral distinctions can be learned.

Natural-language calibration adds a second problem: how should persona text itself be parameterized and searched? We introduce an anchored representation that preserves the authored prompt at zero adjustment while exposing a small set of interpretable preference directions. In an initial frozen cohort spanning Gemma 3 4B, Gemma 3 12B, Qwen 3.5 9B, and Phi-4 14B, anchored calibration improves mean fresh-held-out behavior for every model, but the useful correction is simulator-dependent. Under limited query budgets, structured search based on economic residuals or broad screening transfers more reliably than fixed-order search or a history-aware LLM selector. A further GEPA stress test shows that richer free-text evolution can help some models, but is less reliable in aggregate than the constrained coordinate representation. These results suggest that both *what evidence is observed* and *how the persona is allowed to change* shape behavioral generalization.

Finally, we test the same economic-feedback principle on a public Nutri2Cycle discrete-choice benchmark, where latent preference truth is unavailable. Calibration of an initially misspecified structured persona population improves agreement with held-out human choices, although direct estimation on human data remains substantially stronger on the primary held-out share and likelihood metrics. EIPG should therefore be viewed as a way to use limited behavioral evidence to build more defensible synthetic populations, not as a substitute for observed human data.

Our contributions are fourfold. First, we formulate persona grounding as behavioral calibration and give a unified EIPG framework for structured and natural-language generators. Second, using a controlled benchmark, we show that intervention informativeness, rather than inner-model capacity alone, governs whether calibration transfers to unseen counterfactuals. Third, we introduce an anchored natural-language parameterization and show that useful corrections are simulator-dependent, while structured allocation of finite behavioral-query budgets improves search reliability. Fourth, on a public human discrete-choice benchmark, we show that calibrating an initially misspecified persona population improves transfer to untouched outcomes and unseen respondents, while direct human estimation remains substantially stronger on held-out Task 4 choice-share error and negative log-likelihood.

2 Related Work

Persona-conditioned simulators and grounding. LLMs are increasingly used as synthetic respondents, economic agents, and user simulators for surveys, recommendation, and market research (Horton et al., 2023; Argyle et al., 2023; Aher et al., 2023; Brand et al., 2023; Park et al., 2023; Zhang et al., 2024). Recent work also treats personas as first-class representations, from generated synthetic populations (Paglieri et al., 2026) to behavior-derived interest personas (Wang et al., 2026). This raises a validity problem: fluent or persona-consistent responses need not reproduce the tradeoffs needed for downstream decisions. PersonaGym evaluates behavioral consistency with assigned personas (Samuel et al., 2025), while other work documents bias, instability, or gaps between generated personas and real users (Liu et al., 2024; Gupta et al., 2024; Li et al., 2025a; Meshi et al., 2026). Complementary work grounds simulators or personalized agents in authentic traces, dense respondent histories, human feedback, or limited human samples (Toubia et al., 2025; Zhu et al., 2026; Liang et al., 2026; Zhao et al., 2025; Krsteski et al., 2026; Wang et al., 2025; Zhang et al., 2026). These approaches typically use user-specific evidence to improve response prediction or estimation. EIPG instead treats the generator as the object of calibration, aiming to construct a reusable population whose induced behavior remains credible under intervention.

Preference elicitation, choice modeling, and identification. Classical and active elicitation methods seek informative queries that reveal latent preferences (Chen & Pu, 2004; Vayanos et al., 2020), while recent work elicits preferences through language models (Li et al., 2025b; Soumalias et al., 2025). Discrete-choice models provide a complementary language for substitution, elasticities, willingness-to-pay, and counterfactual responses (McFadden, 1980; Berry et al., 1995; Train, 2009). EIPG uses these ideas at the population-generator level: interventions are valuable insofar as they distinguish persona populations that fit similar observed choices but imply different behavior out of sample. This shifts emphasis from in-sample fit to preservation of behaviorally relevant substitution patterns under unseen interventions.

Persona control and prompt optimization. Persona vectors and activation-space methods steer behavioral traits inside an LLM (Chen et al., 2025; Izawa et al., 2026; Pai et al., 2026), while methods such as GEPA optimize textual instructions using reflective feedback (Agrawal et al., 2026). EIPG instead optimizes agreement with behavioral moments and evaluates transfer to frozen contexts. This motivates our anchored text representation: interpretable preference directions are adjustable while the authored persona remains the reference point.

3 METHOD

3.1 PERSONA-GENERATOR CALIBRATION

EIPG treats persona construction as part of the simulator rather than as fixed preprocessing. For a choice context $\mathbf{x} \in \mathcal{X}$ and categorical outcome $y \in \mathcal{Y}$, a persona generator G_ϕ induces a population over persona representations $\mathbf{z}$, and a behavioral simulator produces $y \sim \pi_\theta(y \mid \mathbf{x}, \mathbf{z})$. The representation $\mathbf{z}$ may be a structured preference vector or a natural-language persona prompt. In either case, the goal is to choose generator parameters ϕ so that the *behavior induced by the resulting population* matches reference behavioral evidence.

A candidate generator produces a population; the simulator turns personas and contexts into choices; those choices are summarized into behavioral moments; and discrepancies from target moments become feedback for the generator. Figure 1 summarizes this common pipeline. We distinguish simulation contexts $\mathcal{X}_{\text{sim}}$, anchor contexts $\mathcal{X}_H$, calibration interventions $\mathcal{X}_{\text{calib_int}}$, and held-out counterfactual contexts $\mathcal{X}_{\text{cf}}$. The last set never enters candidate-level optimization, so low calibration error cannot be mistaken for evidence of counterfactual validity.

We use bold Latin symbols for vector- or matrix-valued objects, while Greek parameters use standard math font with dimensions stated explicitly. Thus $\mathbf{w} = (w_1, \ldots, w_{K_z})$ and w_k is scalar; likewise, M_r^{tar} is the scalar rth entry of $\mathbf{M}_{\text{tar}}$. For the structured regime, the generator is a mixture over economic preference vectors,

$$p_\phi(\mathbf{z}) = \sum_{k=1}^{K_z} w_k \mathcal{N}(\mathbf{z}; \mu_k, \Sigma_k), \qquad \mathbf{z} \in \mathbb{R}^{d_z},\ \mu_k \in \mathbb{R}^{d_z},\ \Sigma_k \in \mathbb{R}^{d_z \times d_z}, \tag{1}$$

with learnable mixture weights, component means, and covariance matrices. The natural-language regime uses the same abstraction but exposes sparse adjustments around authored prompts (Section 3.4). In both regimes, generator parameters are selected through behavior rather than direct persona labels.

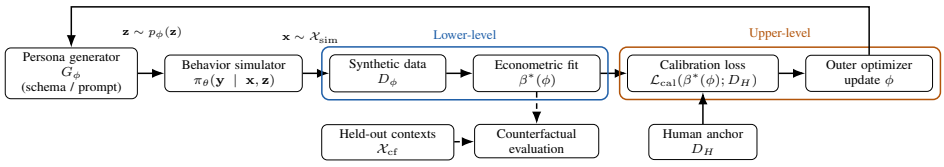


Figure 1: **EIPG calibration.** A persona generator induces a synthetic population, the simulator produces behavior, and behavioral discrepancies update the generator. Structured variants estimate an intermediate economic model; natural-language variants compare behavioral moments directly.

3.2 BILEVEL ECONOMIC FEEDBACK

The interaction between structured data generation and downstream econometric evaluation forms a closed-loop system. We write this dependency as the bilevel program

$$\phi^* \in \arg\min_{\phi \in \Phi} \{\mathcal{L}_{\text{cal}}(\phi) + \lambda \mathcal{R}(\phi)\} \qquad \text{(Upper level: calibration)} \tag{2}$$

$$\text{s.t.} \quad \beta^*(\phi) \in \arg\min_{\beta \in \mathbb{R}^{d_f}} \mathcal{L}_{\text{econ}}(\beta; D_\phi) \qquad \text{(Lower level: economic fit).} \tag{3}$$

Here, $\mathcal{R}(\phi)$ is an optional diversity regularizer and $\lambda \geq 0$ its weight. The upper level searches over persona generators, while the lower level maps the simulated choices from each candidate generator into a common economic summary.

Lower-level objective. For structured generators, sampling and simulating a candidate G_ϕ on $\mathcal{X}_{\text{sim}}$ produces a choice dataset D_ϕ. The inner model is fit by minimizing negative log-likelihood,

$$\mathcal{L}_{\text{econ}}(\beta; D_\phi) = -\sum_{i \in D_\phi} \log p_\beta(y_i \mid \mathbf{x}_i), \tag{4}$$

where the main specification is multinomial logit,

$$p_\beta(y_i = j \mid \mathbf{x}_i) = \frac{\exp(\beta^\top f(\mathbf{x}_i, j))}{\sum_{k \in \mathcal{C}(\mathbf{x}_i)} \exp(\beta^\top f(\mathbf{x}_i, k))}, \tag{5}$$

with behavioral parameter $\beta \in \mathbb{R}^{d_f}$ and per-alternative feature map $f(\mathbf{x}, j) \in \mathbb{R}^{d_f}$. The inner model is an interpretable comparison interface, not an assumption that the latent population has been recovered exactly. For example, two generators may induce similar baseline shares but different price sensitivity; their fitted behavioral summaries can still expose that distinction.

Upper-level objective. Reference data—human where available—define $\mathbf{M}_{\text{tar}} = (M_1^{\text{tar}}, \ldots, M_R^{\text{tar}})^\top$, where M_r^{tar} is the rth target moment. Candidate behavior induces $\mathbf{M}_\phi$, which in the structured regime may be computed from $\beta^*(\phi)$, and the outer loop minimizes

$$\mathcal{L}_{\text{cal}}(\phi) = (\mathbf{M}_\phi - \mathbf{M}_{\text{tar}})^\top \mathbf{W} (\mathbf{M}_\phi - \mathbf{M}_{\text{tar}}). \tag{6}$$

Here, $\mathbf{W}$ is the weighting matrix over behavioral moments. In all structured experiments reported in this paper we use uniform diagonal weighting, $\mathbf{W} = \mathbf{I}$; the matrix notation is retained to make explicit where alternative moment weights would enter. This $\mathbf{W}$ is unrelated to the Gaussian-mixture weights w_k in Equation 1. The feedback is therefore behavioral rather than textual. If the simulated population systematically chooses lower-quality products than the target population, the residual signals a mismatch in the induced quality tradeoff; the optimizer may shift mixture parameters or an exposed language coordinate without ever being given a prescribed persona description.

Natural-language calibration uses the same loop but evaluates behavioral moments directly rather than refitting an econometric model at every prompt update:

persona generator $\rightarrow$ simulated behavior $\rightarrow$ behavioral evidence $\rightarrow$ generator update.

3.3 MOMENTS, INTERVENTIONS, AND IDENTIFICATION

Calibration combines level choice shares, chosen-attribute summaries, substitution patterns, and intervention-response moments. For a generic behavioral statistic m, an intervention response can be written as

$$\Delta m = m(\mathbf{x}_{\text{int}}) - m(\mathbf{x}_{\text{anchor}}),$$

which distinguishes matching a level from matching how behavior changes when the environment changes.

This is the key identification intuition. Suppose two persona populations generate nearly identical baseline shares, but one contains more price-sensitive users and the other more quality-sensitive users. Baseline behavior may not distinguish them even though their counterfactual responses differ. A price intervention separates them along one dimension; a quality intervention adds another restriction; sustainability and novelty interventions can expose further tradeoffs. EIPG therefore treats intervention design as part of the information available for calibration. Weak coverage leaves directions in persona space weakly constrained, whereas broader interventions reduce the set of behaviorally equivalent candidates.

3.4 ANCHORED NATURAL-LANGUAGE PERSONA CALIBRATION

Unrestricted prompt rewriting creates a large and semantically entangled search space. We instead anchor calibration to the authored persona. Let $\mathbf{p}_k^{(0)}$ be the original prompt for persona k and let $\delta_k \in \mathbb{R}^d$ be a sparse vector of relative preference adjustments. A deterministic renderer

$$p_k(\delta_k) = R\Big(\mathbf{p}_k^{(0)}, \delta_k\Big), \qquad R\Big(\mathbf{p}_k^{(0)}, 0\Big) = \mathbf{p}_k^{(0)}$$

returns the original text exactly at zero adjustment and realizes nonzero coordinates as minimal relative clauses.

For example, our quality-oriented persona balances price, quality, brand familiarity, sustainability, and new features. A negative brand adjustment does not rewrite the persona wholesale: the renderer appends the instruction to place slightly less weight on familiar brands and to preserve all other preferences and tradeoffs from the baseline. Each search action therefore has a behavioral interpretation while unrelated parts of the persona remain anchored.

For the resulting prompt population $P(\delta)$, the LLM simulator produces choices on a frozen calibration ensemble and we optimize the weighted moment RMSE used by the implementation,

$$J(\delta) = \sqrt{\frac{1}{R}\sum_{r=1}^{R} \left[a_r \left(M_r(P(\delta)) - M_r^{\mathrm{tar}}\right)\right]^2}.$$

Here r indexes the R moment components. The fixed scaling weights are $a_r = 0.5$ for price chosen-attribute moments and $a_r = 1$ for quality, sustainability, novelty, brand, and intervention-response moments. We use a_r rather than w_r to distinguish these moment-scaling constants from the mixture weights w_k in Equation 1. The base search uses bounded sparse coordinate moves with strict-improvement acceptance. Free-form rewrites and richer text-search methods are evaluated separately rather than being treated as equivalent to this constrained representation.

3.5 OUTER OPTIMIZATION, QUERY BUDGETS, AND FROZEN EVALUATION

The feedback objective does not require a single outer optimizer. Structured generators use black-box evolutionary search; anchored language calibration can use coordinate evaluation, staged screening, or adaptive action selection. Under a finite LLM-query budget, screen–confirm first evaluates all moves on a shared subset of calibration contexts and then spends the remaining budget on the most promising candidates. Adaptive policies instead choose subsequent actions from residuals and previously observed candidate outcomes.

Across all regimes, the candidate-to-behavior-to-residual interface remains fixed, and held-out contexts remain outside candidate selection. Once a strategy is specified, it is evaluated on contexts unavailable during its selection. This guards against repeatedly adapting a persona generator to the same counterfactual set until it appears to generalize. Full optimizer settings, seed mappings, query accounting, and freeze manifests are provided in the appendix.

4 EXPERIMENTS

4.1 EVALUATION REGIMES

We use three regimes with complementary validity roles. The **controlled benchmark** exposes latent and counterfactual truth to isolate identification failures; the **natural-language benchmark** tests the same feedback principle when personas are prompts and behavior is mediated by an LLM; and the **public benchmark** tests transfer to observed human outcomes without access to latent truth. Full construction, optimizer, seed, and freeze details are in the appendix.

4.2 CONTROLLED SYNTHETIC BENCHMARK

Choice contexts contain five alternatives with price, quality, sustainability, novelty, and brand features. Human truth is a three-segment latent population, and the candidate generator has three

mixture components. We compare price-only calibration with interventions spanning price, quality, sustainability, and novelty. Because latent and counterfactual truth are known, this regime separates weak intervention information from insufficient inner-model capacity. CF-dev is used only to diagnose and freeze the calibration design; a separately generated CF-test is revealed afterward.

We report the calibration loss in Equation 6 and counterfactual share error $\mathrm{CFErr} = \|\mathbf{S}_{\phi}^{\mathrm{cf}} - \mathbf{S}_{\mathrm{tar}}^{\mathrm{cf}}\|_2$. Homogeneous MNL, panel latent-class, and oracle variants serve as diagnostics; the frozen final comparison is initial/static versus price-only and all-four EIPG.

4.3 Natural-language persona calibration with local LLMs

We use budget-sensitive, quality-oriented, and sustainability-oriented authored personas over the same five attributes. The anchored representation is calibrated from chosen-attribute and intervention-response moments over frozen context variants. The reference Gemma 3 12B study tests seed robustness and paired held-out transfer; additional free-form, structured-proposal, and local-optimality diagnostics are reported in Appendix D.

The cross-model study holds calibration evidence, decoding, coordinate space, and candidate order fixed across Gemma 3 4B, Gemma 3 12B, Qwen 3.5 9B, and Phi-4 14B. A full one-sweep search evaluates the authored population plus 30 signed coordinate moves, costing 3,720 logical simulator calls per model. Under a 1,320-call budget we compare fixed-prefix, random-subset, screen–confirm, residual-prior LinUCB, and history-aware LLM selection. GEPA 0.1.4 (Agrawal et al., 2026) provides a richer-representation stress test by evolving free-text suffixes attached to immutable authored prompts. Because it changes both representation and search, we do not interpret it as a pure optimizer comparison. Every selected strategy is frozen before its corresponding fresh held-out cohort is evaluated.

4.4 Public choice benchmark

We use the Nutri2Cycle labelled discrete-choice data (Ornelas Herrera & Kallas, 2023), freezing a Belgium-bread protocol with 772 complete respondents. Task 2 is the anchor, Tasks 1 and 3 provide calibration interventions, and Task 4 is an untouched holdout whose organic premium lies outside the Tasks 1–3 range. We compare static personas, anchor-only EIPG, and intervention-rich EIPG; a labelled-alternative MNL fitted directly to human data provides a reference. Primary Task 4 metrics are choice-share L2, held-out human-choice NLL under the simulated-population model, and anchor-to-Task-4 response-vector error. A predeclared respondent holdout on Tasks 1–3 provides a secondary robustness check.

5 Results

5.1 Controlled benchmark: intervention coverage governs transfer

The controlled benchmark first exposes a calibration–generalization gap. With a homogeneous MNL inner model, calibration L2 falls from 0.1211 to 0.1071 while CF-dev L2 rises from 0.0612 to 0.0966. Replacing that summary with a panel latent-class model produces the same qualitative failure and nearly identical candidate rankings. Figure 2a makes the mismatch visible: optimizing a weak calibration signal can move the generator in the wrong counterfactual direction even when the inner model is made more expressive.

The failure changes when the calibration interventions expose more preference dimensions. Expanding feedback from price alone to price, quality, sustainability, and novelty increases mean Pearson-/Spearman calibration–CF correlation in the fixed candidate bank from 0.729/0.646 to 0.788/0.730, while top-decile calibrated candidates have lower CF-dev error. In full-search confirmation, the four-intervention objective improves CF-dev by 14.0% on average versus 4.7% for price-only calibration, and the fraction of runs that worsen CF-dev falls from 36% to 20%. Importantly, the richer design achieves a slightly smaller mean calibration-objective improvement, so the gain is not explained by simply optimizing the same signal more aggressively.

After using CF-dev only to freeze the method-level specification, we generate a separate final counterfactual test. Mean CF-test L2 falls from 0.0704 for the initial generator to 0.0682 for price-only

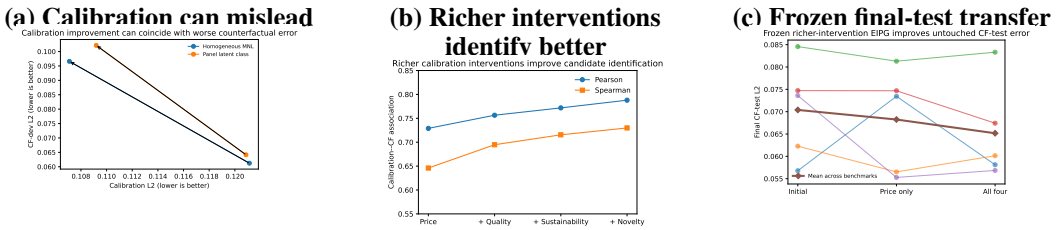


Figure 2: **Controlled benchmark from failure diagnosis to final transfer.** (a) The initially selected generators reduce calibration error while increasing CF-dev error under both homogeneous and latent-class inner models. (b) Broadening intervention coverage strengthens the relationship between calibration quality and counterfactual quality. (c) After freezing the richer calibration design, the all-four specification improves final counterfactual error in four of five independently regenerated benchmarks after averaging nested search seeds. CF-dev is development-only; panel (c) uses the untouched final test.

and 0.0652 for all-four EIPG, corresponding to mean reductions of 1.1% and 6.6%, respectively. The all-four specification improves over the initial generator in four of five fresh benchmark realizations after averaging nested search seeds (Figure 2c).

Full aggregate and benchmark-level final-test tables are reported in Appendix C.5.

Result 1. Informative intervention coverage improves candidate ranking and frozen transfer; a richer latent-class inner model does not repair a weak outer signal.

5.2 Anchored language calibration transfers across contexts and models

On Gemma 3 12B, the anchored parameterization improves all four calibration-seed replications (mean 26.5%). After prompt freeze, all four paired held-out evaluations also improve, with a mean/median gain of 16.7%/10.7% and a range from 3.2% to 42.2%. The accepted coordinates vary across seeds, and free-form rewriting, LLM-proposed sparse updates, and exhaustive local perturbations do not improve the anchored reference solution (Appendix D). The mechanism therefore transfers across fresh contexts even though neither the exact edit nor the effect size is universal.

Holding calibration evidence and a 3,720-call one-sweep budget fixed across Gemma 3 4B, Gemma 3 12B, Qwen 3.5 9B, and Phi-4 14B, all four models pass the same interface/transfer gate and improve in mean fresh-held-out error in the first frozen cross-model cohort (Table 1). The useful correction is strongly simulator-dependent: Gemma 3 4B selects `quality.brand` $+0.5$, Gemma 3 12B selects the opposite `quality.brand` -0.5, Qwen 3.5 9B selects `sustain.brand` -0.5, and Phi-4 14B selects `quality.brand` $+0.5$. Thus the same authored persona text does not define the same behavioral population across simulators.

Model	Selected anchored edit	Mean held-out gain	Improved pairs
Gemma 3 4B	`quality.brand` $+0.5$	22.7%	4/4
Gemma 3 12B	`quality.brand` -0.5	14.8%	2/4
Qwen 3.5 9B	`sustain.brand` -0.5	45.6%	4/4
Phi-4 14B	`quality.brand` $+0.5$	12.0%	3/4

Table 1: **Full-budget cross-model transfer (3 720 calls).** Each row evaluates a prompt population frozen from the common calibration protocol on four fresh held-out contexts. The selected edit itself changes with the simulator, including opposite brand corrections for the two Gemma scales.

Result 2a. Useful corrections are simulator-specific: even Gemma 3 4B and Gemma 3 12B select opposite `quality.brand` adjustments from the same authored persona.

The later sequentially frozen budget studies ask whether these gains require exhaustive coordinate search. On the fifth untouched cohort, residual-prior LinUCB and screen–confirm achieve mean held-out gains of 19.6% and 18.4% at 1,320 calls, improving 12/16 and 14/16 model–context pairs.

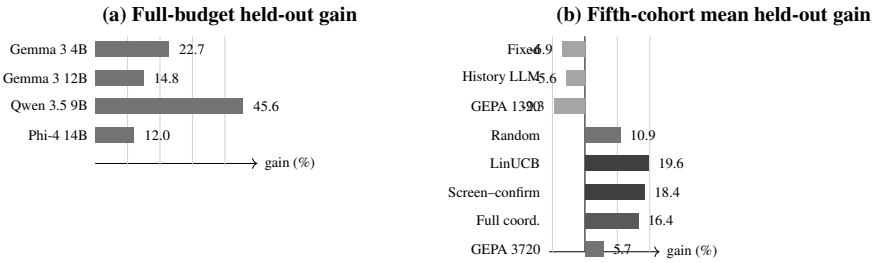


Figure 3: **Natural-language calibration is model-dependent, and search allocation matters.** Left: mean held-out gain from the frozen 3,720-call coordinate sweep on the first cross-model cohort. Right: mean held-out gain on the fifth untouched cohort for the final search-strategy stress test. Positive values denote lower weighted held-out RMSE than the authored-prompt baseline. The panels use different predeclared held-out cohorts and summarize cross-model transfer and query/search robustness.

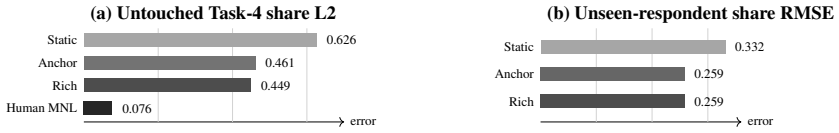


Figure 4: **Public-data transfer at two levels.** Left: primary Task-4 choice-share error after selecting persona generators on Tasks 1–3. Right: pooled task-share RMSE on the predeclared 20% respondent holdout from Tasks 1–3. Calibration improves both evaluations relative to the static generator, while the direct-human MNL remains substantially stronger on the untouched task.

Five random coordinate subsets average 10.9%, while fixed-prefix and history-aware LLM search degrade held-out performance. Full coordinate search at 3,720 calls gains 16.4% and improves 13/16 pairs. GEPA, which evolves free-text suffixes rather than fixed economic coordinates, yields -9.3% at 1,320 calls and +5.7% at 3,720 calls. Its larger-budget benefit is concentrated in Qwen 3.5 9B (+19.5%) and Gemma 3 4B (+3.2%); the frozen selection is the authored baseline for Gemma 3 12B and Phi-4 14B.

The random baseline also reveals why we avoid interpreting any one optimizer as uniquely necessary. Several random subsets recover the same strong Qwen coordinate found by full search, LinUCB, and screen–confirm, while other subsets perform substantially worse. The structured methods therefore appear most useful as systematic ways to reduce search-order and subset sensitivity, not as evidence that a particular policy is universally optimal.

Result 2b. Structured allocation reduces budget sensitivity, while GEPA shows that richer text search can help particular simulators without being more reliable across the cohort.

5.3 Public choice benchmark

The frozen Nutri2Cycle Belgium-bread experiment evaluates persona candidates selected from Tasks 1–3 on untouched Task 4. This removes access to latent population truth and asks the practical question: does calibration selected on limited human behavior improve predictions of a genuinely unseen human choice task?

As summarized in Figure 4, on Task 4 calibration reduces share L2 from 0.6263 to 0.4609 (anchor-only) and 0.4491 (intervention-rich), with corresponding NLL improvement for the richer specification from 1.8705 to 1.5674. The richer specification has only a modest advantage on these two metrics, but its anchor-to-Task-4 response-vector error falls from 0.0348 to 0.0181 relative to anchor-only. The direct-human MNL remains far stronger on share L2 (0.0758) and NLL (1.1811). On the predeclared respondent holdout (Appendix E), both calibrated generators reduce pooled task-share RMSE by about 22%, but intervention-response RMSE does not improve and anchor-only and intervention-rich are indistinguishable.

Result 3. Calibration transfers to an untouched human choice task and unseen respondents, but does not replace direct human estimation; richer intervention moments add only task-dependent gains over anchor-only calibration.

6 Discussion and Conclusion

Across the three regimes, the evidence supports a narrower claim than "better persona prompting": grounding persona simulators is an identification problem. In the controlled benchmark, weak calibration evidence can move the generator away from correct counterfactual behavior, and a richer inner model does not repair that failure; broader interventions help by constraining more behaviorally relevant directions. The public benchmark supplies a weaker external counterpart: calibration transfers to an untouched human task and unseen respondents, but intervention-rich calibration adds only modest Task 4 gains and no respondent-holdout advantage over anchor-only calibration. More moments are useful only when they add information about the behavior of interest: in the public benchmark, the richer objective helps most clearly on the held-out response-vector metric rather than uniformly across all diagnostics. Because Nutri2Cycle varies only price across calibration tasks, it cannot reproduce the multidimensional identification test available in the controlled benchmark.

The natural-language experiments show that the deployed simulator is part of the calibrated object. The same authored persona induces different useful corrections across LLMs, including opposite adjustments to the same coordinate. In the first frozen cross-model cohort, the common economic-feedback mechanism improves mean fresh-held-out behavior for all four tested simulators, while later cohorts show that effect size and reliability remain context-dependent. A persona library should therefore not be assumed to transfer unchanged when the simulator changes; model replacement should trigger behavioral revalidation even when persona text is fixed. Representation and query allocation also matter under finite evidence: screen–confirm and residual-prior LinUCB reduce budget sensitivity, whereas random subsets are variable. GEPA can discover transferable free-text edits for some simulators, especially Qwen 3.5 9B at larger budget, but because it changes both representation and search, it is a stress test rather than a pure optimizer ablation. The broader lesson is that interpretable constraints can regularize calibration without implying one universal edit or optimizer.

EIPG is not a substitute for direct human estimation. On Nutri2Cycle, the direct-human MNL remains substantially stronger on the primary Task 4 share-L2 and NLL metrics. Rather, EIPG is useful when the downstream object is a reusable simulator or synthetic population and only limited behavioral evidence is available to discipline it. Within the tested domain, held-out improvement shows that calibration can transfer beyond the contexts used for candidate selection; it does not establish a uniquely correct latent population, individual-level psychological fidelity, conversational realism, or validity under arbitrary interventions. Scope is also limited: the language study uses four open local models, three personas, and structured product choice; proprietary models, larger persona populations, and open-ended multi-turn interaction remain untested. The public benchmark covers one product category, one country sample, and four repeated tasks, while the controlled benchmark assumes the relevant preference axes are known. Our behavioral moments are deliberately low-dimensional and interpretable, but may omit behavior that matters downstream; logical query counts also do not directly represent latency, monetary cost, or energy use.

Within these limits, the design principle is simple: treat authored personas as an initial representation rather than behavioral ground truth. Ground them against informative behavioral evidence, restrict adjustable degrees of freedom to what that evidence can support, and evaluate on counterfactual contexts outside the optimization loop. Persona quality should ultimately be judged by whether the induced synthetic population behaves credibly under the changes relevant to the downstream decision problem.

Code for reproducibility at https://anonymous.4open.science/r/gomgom2-AA06/README.md

AI USE STATEMENT

Large language model tools were used during research development for language editing, code assistance and debugging, and workflow support. The authors determined the research questions and experimental protocols, executed and inspected and unit-tested verified experimental outputs, and remain responsible for all claims, analyses, and submitted content.

REPRODUCIBILITY STATEMENT

The paper separates development diagnostics from frozen held-out evaluation in all three experimental regimes. Appendix material documents dataset construction, calibration and counterfactual splits, cross-model cohort freezes, five sequential natural-language held-out cohorts, query-budget and adaptive-search strategies, the frozen GEPA stress test, seed mappings, negative prompt-optimization diagnostics, and respondent-holdout robustness. The accompanying code contains frozen experiment configurations and scripts for rebuilding the reported derived artifacts; the public-data source is identified by DOI, while raw respondent-level data are not redistributed.

REFERENCES

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. Gepa: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0e9e708b6f48e14fd0ac29e167413f76-Abstract-Conference.html.

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371. PMLR, 2023.

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.

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, 1995. doi: 10.2307/2171802.

James Brand, Ayelet Israeli, and Donald Ngwe. Using llms for market research. Technical Report 23-062, Harvard Business School, 2023.

Li Chen and Pearl Pu. Survey of preference elicitation methods. Technical Report IC/2004/67, Swiss Federal Institute of Technology in Lausanne (EPFL), 2004.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias runs deep: Implicit reasoning biases in persona-assigned llms. In *International Conference on Learning Representations*, 2024.

John J. Horton, Apostolos Filippas, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from homo silicus? Technical Report 31122, National Bureau of Economic Research, 2023.

Yoshihiro Izawa, Gouki Minegishi, Koshi Eguchi, Sosuke Hosokawa, and Kenjiro Taura. Steering at the source: Style modulation heads for robust persona control. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=JlO2MCzUto.

Stefan Krsteski, Giuseppe Russo, Serina Chang, Robert West, and Kristina Gligorić. Valid survey simulations with limited human data: The roles of prompting, fine-tuning, and rectification. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10887–10906, 2026. doi: 10.18653/v1/2026.acl-long.498.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. Llm generated persona is a promise with a catch. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-2226. Position Paper Track.

Belinda Z. Li, Alex Tamkin, Noah D. Goodman, and Jacob Andreas. Eliciting human preferences with language models. In *International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c9867d5a22653ce98b02595061e40f12-Abstract-Conference.html.

Kaiqu Liang, Julia Kruk, Shengyi Qian, Xianjun Yang, Shengjie Bi, Yuanshun Yao, Shaoliang Nie, Mingyang Zhang, Lijuan Liu, Jaime Fernández Fisac, Shuyan Zhou, and Saghar Hosseini. Learning personalized agents from human feedback. *arXiv preprint arXiv:2602.16173*, 2026.

Andy Liu, Mona Diab, and Daniel Fried. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9832–9850, 2024. doi: 10.18653/v1/2024.findings-acl.586.

Daniel McFadden. Econometric models for probabilistic choice among products. *The Journal of Business*, 53(3):S13–S29, 1980. doi: 10.1086/296093.

Ofer Meshi, Krisztian Balog, Sally Goldman, Avi Caciularu, Guy Tennenholtz, Jihwan Jeong, Amir Globerson, and Craig Boutilier. Convapparel: A benchmark dataset and validation framework for user simulators in conversational recommenders. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5270–5304, 2026. doi: 10.18653/v1/2026.eacl-long.244.

Selene Ivette Ornelas Herrera and Zein Kallas. European consumers' preference and willingness to pay for food products labelled as obtained by a circular farming system—in relation to environmental attitudes and consumption behaviours, 2023.

Davide Paglieri, Logan Cross, William A Cunningham, Joel Z Leibo, and Alexander Sasha Vezhnevets. Persona generators: Generating diverse synthetic personas at scale. *arXiv preprint arXiv:2602.03545*, 2026.

Tsung-Min Pai, Jui-I Wang, Li-Chun Lu, Shao-Hua Sun, Hung-yi Lee, and Kai-Wei Chang. Billy: Steering large language models via merging persona vectors for creative generation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 7870–7915, 2026. doi: 10.18653/v1/2026.eacl-long.369.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023. doi: 10.1145/3586183.3606763.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R. Narasimhan, and Vishvak Murahari. Personagym: Evaluating persona agents and llms. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6999–7022, 2025. doi: 10.18653/v1/2025.findings-emnlp.368.

Ermis Soumalias, Yanchen Jiang, Kehang Zhu, Michael Curry, Sven Seuken, and David C Parkes. Llm-powered preference elicitation in combinatorial assignment. *arXiv preprint arXiv:2502.10308*, 2025.

Olivier Toubia, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen. Twin-2k-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions. *arXiv preprint arXiv:2505.17479*, 2025.

Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

Phebe Vayanos, Yingxiao Ye, Duncan McElfresh, John Dickerson, and Eric Rice. Robust active preference elicitation. *arXiv preprint arXiv:2003.01899*, 2020.

Haoting Wang, Haokai Lu, Zheyun Feng, Jenny Huang, Yifat Amir, Gregory Hinkson, Ben Most, Zelong Zhao, Yixin Kelly Cui, Rein Zhang, Fabio Soldo, Yu Xia, Nihar Bhupalam, Minmin Chen, Konstantina Christakopoulou, Lichan Hong, and Ed H. Chi. Llm-based user personas for recommendations at scale. *arXiv preprint arXiv:2606.12198*, 2026.

Lei Wang, Zikun Ye, and Jinglong Zhao. Efficient inference using large language models with limited human data: Fine-tuning then rectification. *arXiv preprint arXiv:2511.19486*, 2025.

An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. On generative agents in recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1807–1817, 2024. doi: 10.1145/3626772.3657844.

Bohan Zhang, Jiaxuan Li, Ali Hortaçsu, Xiaoyang Ye, Victor Chernozhukov, Anqi Ni, and Edward W. Huang. Agentic economic modeling. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics: Industry Track*, pp. 1136–1149, 2026. doi: 10.18653/v1/2026.acl-industry.80.

Weixiang Zhao, Xingyu Sui, Yulin Hu, Jiahe Guo, Haixiao Liu, Biye Li, Yanyan Zhao, Bing Qin, and Ting Liu. Teaching language models to evolve with users: Dynamic profile modeling for personalized alignment. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1115.

Ming Zhu, Juntao Tan, Rithesh Murthy, Jielin Qiu, Liangwei Yang, Wenting Zhao, Silvio Savarese, Shelby Heinecke, and Huan Wang. Realusersim: Bridging the reality gap in agent benchmarking via grounded user simulation. *arXiv preprint arXiv:2605.20204*, 2026.

A Controlled Synthetic Dataset Construction

This appendix describes the construction of the controlled synthetic datasets used in the symbolic EIPG experiments. The goal of this stage is to instantiate the objects introduced in the main text—persona latents, choice contexts, simulated choices, anchor data, calibration interventions, and held-out counterfactual data—before any econometric model is fit or any generator update is performed.

Run configuration. Table 2 summarizes the configuration used for the run. The configuration fixes the random seed, context sizes, latent features, number of mixture components, number of sampled personas, and observation counts.

Quantity	Value
Mixture components	3
Latent features	price, quality, sustain, novelty, brand
Simulation contexts	300
Anchor contexts	120
Calibration-intervention contexts	120
Held-out counterfactual contexts	120
Alternatives per context	5
Sampled generator personas	160
Synthetic-human personas	320
Initial synthetic observations	1200
Anchor observations	600
Calibration observations	600
Held-out observations	600
Outer budget	24
Outer population	6

Table 2: Run configuration used for controlled synthetic dataset construction.

Generator and persona latents. The initial generator is a mixture distribution over economic preference vectors,

$$p_{\phi_0}(z) = \sum_{k=1}^{K_z} w_k \mathcal{N}(z; \mu_k, \Sigma_k).$$

Table 3 summarizes the sampled segment composition and average latent coordinates for both the initial generator and the synthetic-human truth population. Figure 5 visualizes the sampled initial persona latents along two economically interpretable dimensions.

Source	Segment	Label	Count	Share	mean price	mean quality	mean sustain	mean novelty	mean brand
initial generator	0	budget_sensitive	62	0.39	-1.98	0.98	0.15	0.14	-0.01
initial generator	1	quality_brand_oriented	50	0.31	-0.66	1.83	0.30	0.11	0.95
initial generator	2	sustainability_novelty_oriented	48	0.30	-0.80	0.83	1.63	1.20	0.16
truth population	0	budget_sensitive	133	0.42	-2.33	0.77	0.09	0.01	-0.13
truth population	1	quality_brand_oriented	133	0.42	-0.50	1.98	0.19	-0.04	1.10
truth population	2	sustainability_novelty_oriented	54	0.17	-0.89	0.70	1.78	1.25	0.10

Table 3: Segment composition and average latent coordinates for initial and truth populations.

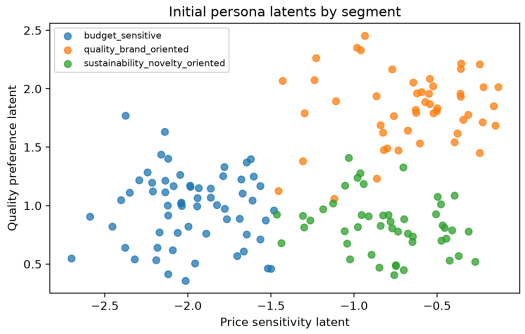


Figure 5: Initial persona latents sampled from G_{ϕ_0}, projected onto price sensitivity and quality preference dimensions.

Context sets. The benchmark constructs four context sets: simulation contexts $\mathcal{X}_{\text{sim}}$, anchor contexts $\mathcal{X}_H$, calibration-intervention contexts $\mathcal{X}_{\text{calib_int}}$, and held-out counterfactual contexts $\mathcal{X}_{\text{cf}}$. Table 4 reports the realized context counts and roles. Table 5 shows one concrete choice context, making the abstract context x visible as a slate of alternatives with observed attributes.

Context set	Role	Contexts	Rows	Alternatives/context	Used for generator updates?
X_sim	Simulation contexts for candidate generators	300	1500	5	yes
X_H	Anchor contexts defining D_H	120	600	5	yes
X_calib_int	Calibration-intervention contexts	120	600	5	yes
X_cf_base	Unperturbed base for held-out interventions	120	600	5	no
X_cf	Held-out counterfactual contexts	120	600	5	no

Table 4: Context sets constructed for the controlled synthetic benchmark.

Context	Alt.	Product	Price	Quality	Sustain.	Novelty	Brand
sim_0000	0	item_00	1.52	0.81	0.51	0.63	0.00
sim_0000	1	item_01	0.58	0.92	0.24	0.93	0.00
sim_0000	2	item_02	0.83	0.28	0.82	0.23	0.00
sim_0000	3	item_03	0.78	0.82	0.21	0.80	1.00
sim_0000	4	item_04	0.76	0.89	0.74	0.52	0.00

Table 5: Example anchor choice context represented as a slate of alternatives.

Choice generation. Given a persona latent z and context x, the symbolic simulator assigns utility $U(j \mid x, z) = z^\top f(x, j)$ to each alternative and samples a choice from the induced logit probabilities. Figure 7 summarizes how the initial generator, truth population, context sets, and symbolic simulator produce the four choice datasets. Table 6 reports the size and role of each generated choice dataset, while Table 7 shows example chosen records.

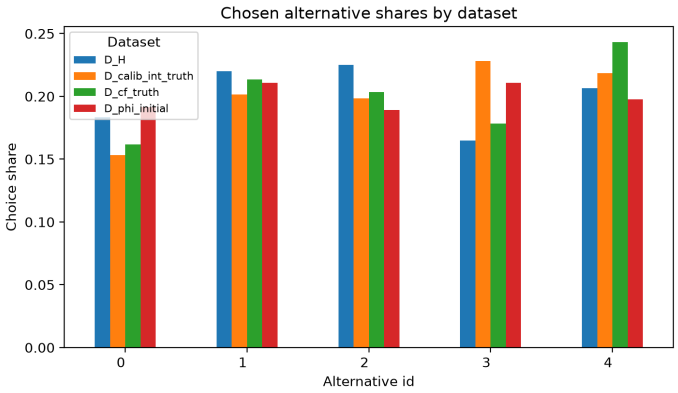


Figure 6: Chosen-alternative shares across generated datasets before MNL fitting or generator optimization.

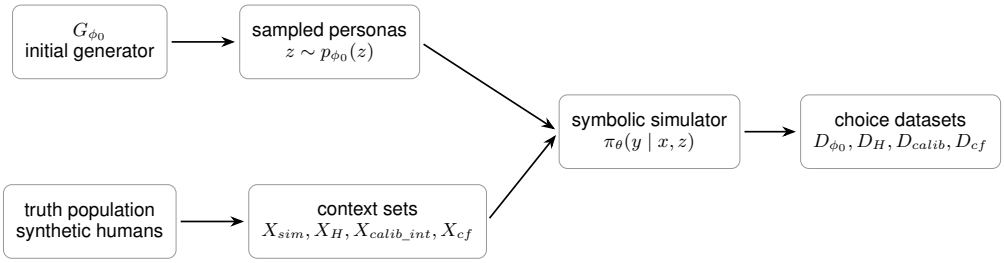


Figure 7: **EIPG bilevel optimization loop.** The generator G_ϕ samples personas from $p_\phi(z)$, which are paired with simulation contexts and queried via simulator π_θ to construct simulated data D_ϕ. The inner loop fits $\beta^*(\phi)$ to D_ϕ; the outer loop updates ϕ to minimize calibration loss relative to human anchor D_H plus a diversity regularizer. Held-out contexts $\mathcal{X}_{\text{cf}}$ are used only for counterfactual evaluation.

Dataset	Role	Obs.	Rows	Contexts	Personas	Used for generator updates?	Chosen rows
D_phi_initial	Initial synthetic data induced by G_phi0	1200	6000	293	160	candidate baseline only	1200
D_H	Anchor behavior from truth population	600	3000	119	278	yes	600
D_calib_int_truth	Calibration-intervention targets	600	3000	120	277	yes	600
D_cf_truth	Held-out counterfactual truth	600	3000	119	280	no	600

Table 6: Generated choice datasets produced by the dataset-construction stage.

Initial diagnostic. Figure 6 compares realized chosen-alternative shares across the initial synthetic dataset, anchor dataset, calibration-intervention truth dataset, and held-out counterfactual truth dataset. This diagnostic is not a calibration result: it is a pre-estimation check that the generated datasets contain meaningful variation across contexts and interventions.

Dataset	Obs.	Context	Persona	Segment	Chosen alt.	Product	Sim. probability
D_phi_initial	D_phi_initial_000000	sim_0003	74	budget_sensitive	3	item_03	0.218
D_phi_initial	D_phi_initial_000001	sim_0223	104	quality_brand_oriented	3	item_03	0.641
D_phi_initial	D_phi_initial_000002	sim_0032	31	sustainability_novelty_oriented	0	item_00	0.487
D_H	D_H_000000	sim_0086	208	budget_sensitive	4	item_04	0.133
D_H	D_H_000001	sim_0107	226	quality_brand_oriented	1	item_01	0.086
D_H	D_H_000002	sim_0098	142	quality_brand_oriented	3	item_03	0.601

Table 7: Example chosen records from the initial synthetic and anchor datasets.

B Inner-Loop Economic Model Estimation

This appendix describes the inner-loop economic estimation stage used in the controlled synthetic benchmark. At this stage, the synthetic dataset D_{ϕ_0} has already been generated by sampling personas from the initial generator and simulating choices with the symbolic random-utility simulator. The role of the inner loop is to fit an econometric summary model to those *simulated* choices; it does not generate the choices itself.

B.1 Estimation Problem

For a fixed generator G_{ϕ_0}, the dataset construction stage produces $D_{\phi_0} = \{(x_i, z_i, y_i)\}_{i=1}^{n_S}$.

The inner loop fits a multinomial logit model to this dataset: $\hat{\beta} = \beta^*(\phi_0) = \arg\min_\beta \mathcal{L}_{\text{econ}}(\beta; D_{\phi_0})$. For each alternative j in context x_i, the fitted model assigns systematic utility $\hat{v}_{ij} = \hat{\beta}^\top f(x_i, j)$, which induces probabilities

$$p_{\hat{\beta}}(y_i = j \mid x_i) = \frac{\exp(\hat{v}_{ij})}{\sum_{k \in \mathcal{C}(x_i)} \exp(\hat{v}_{ik})}.$$

This fitted model is then evaluated on the initial synthetic data, the anchor data, the calibration-intervention data, and the held-out counterfactual data.

Quantity	Value
Run seed	123
Training dataset	D_{ϕ_0}
Training observations	1200
Training rows	6000
Choice model	Multinomial logit
Feature vector	price, quality, sustain, novelty, brand
L2 penalty	0.000
Optimizer success	True
Iterations	8
Training NLL / obs.	1.458

Table 8: Configuration and fit summary for the inner-loop MNL model.

B.2 Fitted Coefficients

Table 9 reports the fitted coefficients. The sign and magnitude of these coefficients summarize the average choice behavior induced by the initial synthetic population and simulator. Figure 8 visualizes the same fitted coefficients.

B.3 Evaluation Across Dataset Splits

After fitting on D_{ϕ_0}, we evaluate the same fitted model on all relevant splits. Table 10 reports negative log-likelihood per observation and top-choice accuracy. These quantities are descriptive diagnostics for the inner model; the later calibration objective uses behavioral moments rather than likelihood alone.

Feature	Interpretation	$\hat{\beta}$	Sign
price	Price sensitivity	-1.160	negative
quality	Quality preference	1.143	positive
sustain	Sustainability preference	0.575	positive
novelty	Novelty preference	0.609	positive
brand	Brand preference	0.366	positive

Table 9: Fitted MNL coefficients from the initial synthetic dataset.

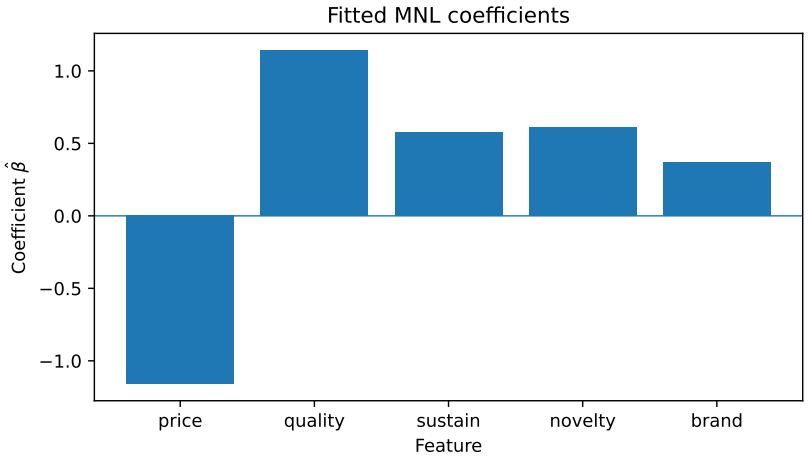


Figure 8: Fitted MNL coefficients for the initial synthetic dataset D_{ϕ_0}.

Dataset	Role	Obs.	Rows	NLL / obs.	Top-choice acc.
D_phi_initial	inner-loop training	1200	6000	1.458	0.363
D_H	anchor behavior	600	3000	1.474	0.360
D_calib_int_truth	calibration intervention target	600	3000	1.450	0.365
D_cf_truth	held-out counterfactual target	600	3000	1.432	0.370

Table 10: Evaluation of the initial fitted MNL model across dataset splits.

Figure 9 compares empirical anchor choice shares with MNL-implied anchor choice shares. The closer the points are to the diagonal, the more closely the fitted model reproduces aggregate anchor shares.

C Controlled Benchmark Diagnostics and Robustness

This section collects the controlled-benchmark evidence that supports the mechanism interpretation in the main text. The main paper reports only the intervention-identification result and the frozen final-test summary; here we show the failure diagnostics, inner-model-capacity comparison, development confirmation, and benchmark-level final-test breakdown.

C.1 Calibration–counterfactual divergence

The earliest outer-loop experiments expose a failure mode that motivates the later intervention analysis. With a homogeneous MNL inner model, the selected generator improves calibration L2 from 0.1211 to 0.1071 while CF-dev L2 worsens from 0.0612 to 0.0966. A matched search using a panel latent-class inner model shows the same qualitative pattern: calibration L2 improves from 0.1209 to 0.1092, while CF-dev L2 worsens from 0.0642 to 0.1021.

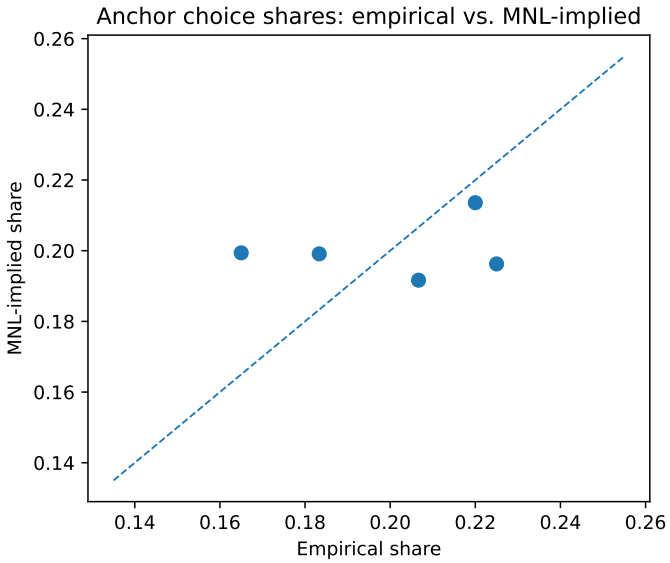


Figure 9: Empirical versus MNL-implied choice shares on the anchor dataset D_H.

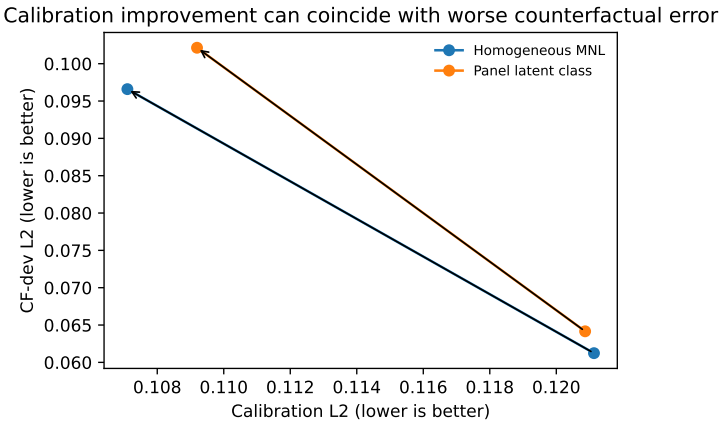


Figure 10: **Calibration improvement can coincide with worse counterfactual error.** Initial and selected generators from matched outer-loop searches move toward lower calibration error but higher CF-dev error. CF-dev is excluded from candidate-level selection.

This divergence is the central diagnostic motivating the rest of the controlled study. It shows that a generator can fit the calibration signal more closely without recovering the preference structure required for unseen interventions.

C.2 Inner-model capacity diagnostic

One possible explanation is that the homogeneous MNL inner model compresses heterogeneous simulated behavior too aggressively. We therefore compare it with richer and oracle-style summaries. Direct simulator prediction obtains CF L2 0.0490. An oracle mixture of segment-specific models, marginalized because target latent labels are unavailable at prediction time, obtains 0.0625. The estimated panel latent-class model obtains 0.0636, close to the oracle-mixture benchmark, while homogeneous MNL obtains 0.0649.

In the actual matched outer-loop search, the homogeneous and panel latent-class models rank the same candidate generators almost identically: Pearson correlation is approximately 0.993 for

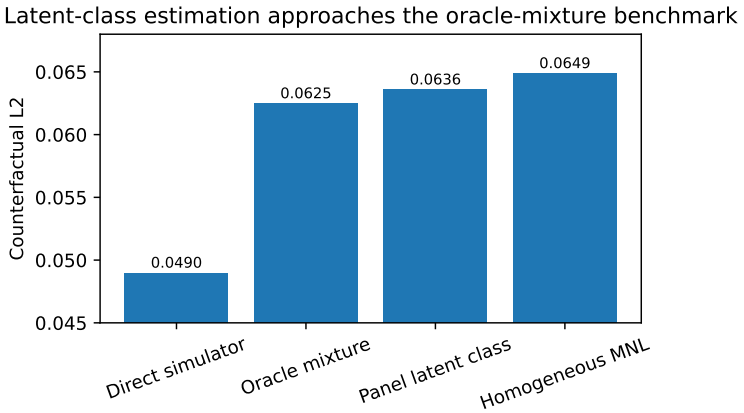


Figure 11: **Inner-model capacity diagnostic.** Counterfactual L2 for direct simulator prediction, an oracle mixture without target labels, the estimated panel latent-class model, and homogeneous MNL.

Inner model	Initial CF-dev l_2	Selected CF-dev l_2
Homogeneous MNL	0.0612	0.0966
Panel latent class	0.0642	0.1021

Table 11: Controlled outer-loop comparison with matched search settings. Increasing inner-model heterogeneity does not remove the calibration–counterfactual divergence in the evaluated search region.

calibration/objective values and 0.996 for CF-dev, with corresponding Spearman correlations of approximately 0.990 and 0.970. They also select the same search path. Thus, in this benchmark and evaluated search region, increasing inner-model heterogeneity alone does not materially repair the outer signal.

The routed segment-specific diagnostic considered during development is deliberately not used as a main benchmark because it relies on privileged target latent labels. It is useful only as an upper-bound mechanism diagnostic.

C.3 Intervention coverage diagnostic

To isolate identification from search dynamics, we score a fixed candidate bank under increasingly rich calibration-intervention bundles: price only; price plus quality; price, quality, and sustainability; and all four continuous attributes after adding novelty. The main text reports the resulting increase in calibration–CF association. Table 12 provides the full bundle-level summary.

The strongest change in the initial benchmark occurs after sustainability variation is added, but this should not be interpreted as a domain-general claim that sustainability is intrinsically more informative. The relevant conclusion is that broadening the intervention span exposes preference dimensions that are otherwise weakly identified by the calibration objective.

C.4 Outer-loop confirmation across benchmark and search seeds

We next compare the price-only and four-intervention objectives in the actual EIPG search. The confirmation experiment uses five independently regenerated benchmark realizations and five search seeds within each benchmark, for 25 search replications per condition. Search seeds within a benchmark share the same benchmark realization and therefore are not treated as independent datasets.

The richer design improves CF-dev by 14.0% on average compared with 4.7% for price-only calibration, while the fraction of runs with worse CF-dev falls from 36% to 20%. The calibration–CF

Calibration interventions	Pearson r	Spearman ρ	Selected CF-dev	Top-10% CF-dev
Price only	0.729	0.646	0.0629	0.0635
Price + quality	0.756	0.695	0.0547	0.0584
Price + quality + sustainability	0.772	0.715	0.0505	0.0584
All four	0.788	0.730	0.0515	0.0566

Table 12: Fixed-candidate-bank identification diagnostic averaged across five independently regenerated controlled benchmarks. Lower CF-dev is better; higher correlation indicates that the calibration objective ranks counterfactually good candidates more reliably.

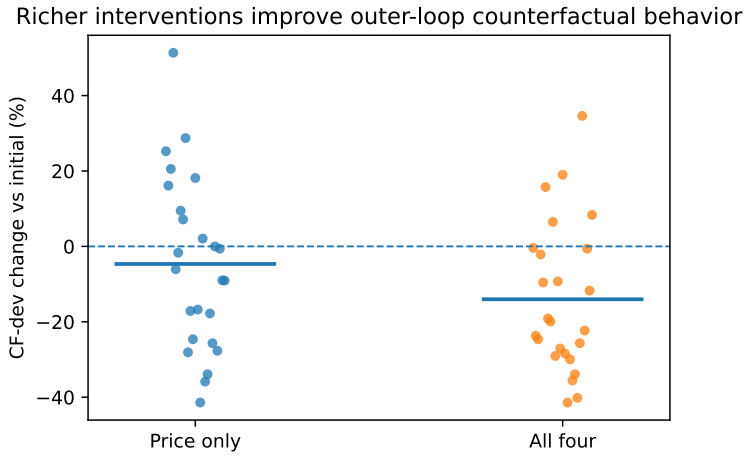


Figure 12: **Development-stage outer-loop confirmation.** CF-dev change relative to the initial generator across 25 matched search replications per calibration design. Negative values indicate improvement; horizontal segments show condition means.

association also strengthens. Importantly, the richer design achieves a slightly smaller mean calibration improvement, which argues against interpreting the result as merely stronger optimization of the same signal.

CF-dev remains a development diagnostic throughout this analysis. It is excluded from candidate-level selection but is used at the meta level to compare and freeze the calibration design.

C.5 Untouched final test by benchmark realization

After freezing the four-intervention equal-weighted specification and the price-only comparator, final counterfactual truth is generated independently. There are five fresh benchmark realizations, each with three predeclared search seeds. Table 15 averages search seeds within benchmark so that the independent benchmark level is visible explicitly.

Method	n	Mean CF-test L2	Median	Mean change vs initial	Improved / n
Initial static	15	0.0704	0.0716	–	–
Price only	15	0.0682	0.0720	-1.1%	10/15
All four	15	0.0652	0.0634	-6.6%	10/15

Table 14: Untouched final counterfactual test for the frozen symbolic protocol. The final test was generated only after the calibration design and search settings were frozen. Negative change denotes lower error than the initial generator.

The four-intervention specification improves over the initial generator in four of five benchmark realizations after averaging over search seeds. The remaining benchmark illustrates the heterogeneity

Calibration design	n	Calib. improvement	CF-dev change	CF worsened	Pearson r	Spearman ρ
Price only	25	14.2%	-4.7%	36%	0.552	0.458
All four	25	13.4%	-14.0%	20%	0.697	0.613

Table 13: Development-stage outer-loop confirmation across five benchmark seeds and five search seeds per benchmark. Negative CF-dev change denotes improvement. CF-dev is used only for meta-level diagnostics and never for candidate-level selection.

Benchmark seed	Initial	Price only	All four
1 701	0.0568	0.0734	0.0581
1 702	0.0623	0.0565	0.0601
1 703	0.0846	0.0813	0.0833
1 704	0.0747	0.0747	0.0674
1 705	0.0736	0.0553	0.0568

Table 15: Untouched final counterfactual l_2 averaged over the three predeclared search seeds within each independent benchmark realization.

emphasized in the main text: the richer design improves counterfactual validity on average, but it does not guarantee improvement on every realization.

C.6 Interpretation and scope

These diagnostics support a narrow mechanism claim. In the controlled benchmark, the original failure is not removed by increasing downstream choice-model capacity, whereas increasing the information content of the calibration interventions improves candidate ranking and changes the behavior of the actual outer loop. The conclusion is local to the evaluated generator family, intervention family, benchmark, and search region. In particular, all controlled interventions act on the designated alternative and on continuous attributes; categorical brand interventions and unrestricted natural-language prompt edits are not part of this experiment.

D Anchored Natural-Language Persona Calibration

This appendix documents the natural-language persona experiment reported in Section 5.2. The goal is to make the textual calibration protocol auditable while preserving the separation between calibration and final held-out evaluation.

D.1 Authored persona prompts and anchored rendering

The experiment starts from three authored persona descriptions. These prompts are intentionally plausible but behaviorally under-grounded: they specify broad tendencies without encoding precise economic tradeoffs. Table 16 reports the exact baseline prompts and the reference-seed calibrated prompt population.

The anchored renderer introduces a structured adjustment vector δ_k for each persona k. At $\delta_k = 0$, the renderer returns the authored prompt byte-for-byte. Nonzero coordinates append a short relative clause while explicitly preserving unrelated preferences and tradeoffs. The reference run uses five preference dimensions: price, quality, sustainability, novelty, and brand.

D.2 Calibration protocol and seed robustness

Each calibration evaluation uses five frozen context variants per intervention and 120 persona–context choice observations. The objective is weighted RMSE over segment-specific chosen-attribute moments and global intervention-response moments. Coordinate search uses step size 0.5, bounds $[-1.5, 1.5]$, and strict-improvement acceptance.

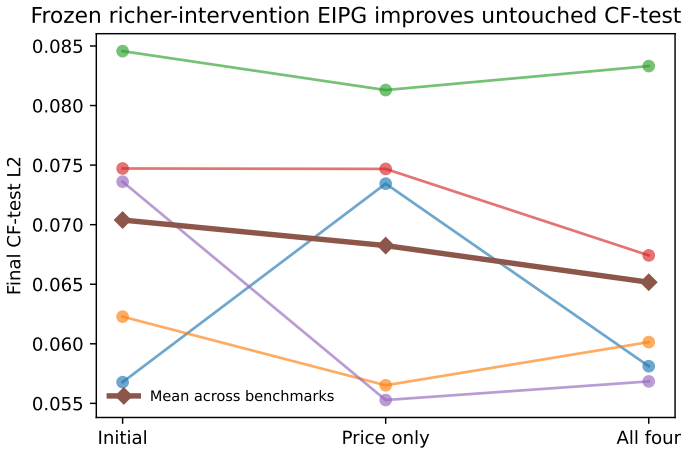


Figure 13: **Untouched final counterfactual test by benchmark realization.** Thin lines show the five independently regenerated final benchmark realizations after averaging over the three nested search seeds. The thick line shows the benchmark mean.

Persona	Authored prompt	Reference calibrated prompt (seed 1729)
Budget	A practical shopper who looks for good overall value. They compare price, quality, and familiar brands, and they are open to sustainability or novelty when those features make the offer more appealing.	Unchanged from authored prompt.
Quality	A balanced shopper who considers price, product quality, and brand familiarity together. They also notice sustainability and new features, but usually weigh several aspects before deciding.	A balanced shopper who considers price, product quality, and brand familiarity together. They also notice sustainability and new features, but usually weigh several aspects before deciding. For this simulation, relative to that baseline description, place slightly less weight on familiar brands. Preserve all other preferences and tradeoffs from the baseline.
Sustainability	An environmentally aware shopper who likes sustainable and innovative products, but usually balances those benefits against price, quality, and brand familiarity rather than paying a strong premium.	Unchanged from authored prompt.

Table 16: **Exact authored natural-language personas and reference calibrated prompts.** Zero anchored adjustment reproduces the authored text exactly.

The reference seed 1729 begins at objective 0.080786 and reaches 0.064187, a 20.55% reduction, after accepting the update $\delta_{\text{quality,brand}} = -0.5$. A subsequent coordinate sweep finds no strict improvement. We then repeat the same frozen mechanism under three additional calibration-context seeds. All four seeds improve relative to their original authored prompts, with mean improvement 26.53%, median improvement 26.05%, minimum 20.55%, and maximum 33.45%.

The identity of the first accepted coordinate varies across seeds, as shown in Table 17. Only one of four seeds begins with the same move as the reference run. We therefore interpret the robustness result as stability of the anchored calibration mechanism rather than stability of a single universal textual edit.

Calibration seed	First accepted move	Match reference?
1729	quality.brand -0.50	yes
2027	quality.quality $+0.50$	no
2718	budget.quality -0.50	no
3141	quality.quality $+0.50$	no

Table 17: **First accepted anchored coordinate by calibration-context seed.** The mechanism improves all four seeds even though the first selected coordinate varies.

Diagnostic	Proposed	Unique evaluated	Best result
Simultaneous free-form rewrite	2 iterations	–	anchored retained
Persona-wise free-form rewrite	3 personas	3	anchored retained
Diverse persona/strategy rewrite	12	12	$0.067028 > 0.064187$
Structured LLM multi-coordinate proposals	8	6	anchored retained

Table 18: **Calibration-only LLM outer-search diagnostics.** More flexible prompt or structured proposal mechanisms do not improve the anchored reference incumbent.

D.3 Outer-search diagnostics

After obtaining the anchored reference solution, we test whether a more flexible LLM-based outer optimizer can improve it. These diagnostics are calibration-only and do not access held-out contexts.

First, simultaneous free-form residual-guided rewriting selects the anchored iteration-0 population as best. Persona-wise rewriting also fails to improve the incumbent. We then generate twelve distinct persona/strategy rewrite proposals spanning minimal, balanced, response-focused, and segment-focused editing strategies. All twelve are unique at the proposal level and none improves the anchored objective. The best new free-form candidate is the sustainability-oriented response-focused rewrite with objective 0.067028, compared with the anchored objective 0.064187.

Next, the LLM is restricted to proposing sparse structured updates in the anchored coordinate space rather than writing prompt text. Eight proposal strategies produce six unique valid multi-coordinate candidates, again with no strict improvement. These negative results motivate a deterministic local-optimality check.

D.4 Exhaustive local-optimality check

To separate poor LLM proposal quality from local stability of the anchored solution, we exhaustively evaluate all single-coordinate and all two-coordinate perturbations with step ± 0.25 around the reference solution. This produces 450 population candidates. Expensive simulator calls are factorized across persona states, requiring 153 unique persona states rather than 450 independent full-population evaluations. No candidate strictly improves the anchored objective. Thus the reference solution is locally optimal over this complete one- and two-coordinate neighborhood at step 0.25.

D.5 Frozen held-out protocol

Only after calibration, robustness analysis, and search diagnostics are complete do we construct held-out contexts. For each calibration seed, the selected prompts are frozen and paired with a predeclared independent held-out context seed. The held-out runner contains no optimization, rewriting, coordinate updates, or prompt-selection path.

Across the four frozen pairs, every calibrated population improves over its original authored-prompt baseline. Held-out improvement ranges from 3.19% to 42.15%, with mean 16.68% and median 10.69%. The spread is substantial, so we interpret the result as evidence of consistent directional transfer rather than a uniform effect size.

Cache behavior. Original held-out evaluations have zero cache reuse. Calibrated populations show cache-hit rates of one-third or two-thirds because some personas remain byte-for-byte identical

Quantity	Value
Reference anchored objective	0.064187
Perturbation step	± 0.25
Population candidates evaluated	450
Unique persona states evaluated	153
Strictly improving candidates	0

Table 19: **Exhaustive local-optimality diagnostic around the reference anchored solution.**

Calibration seed	Held-out seed	Original RMSE	Calibrated RMSE
1729	11729	0.092461	0.083522
2027	12027	0.092081	0.081304
2718	12718	0.098108	0.056759
3141	13141	0.064144	0.062100

Table 20: **Frozen paired held-out evaluation.** Original and calibrated persona populations are evaluated on identical held-out contexts within each row. The corresponding reductions are 9.67%, 11.70%, 42.15%, and 3.19%.

to their authored prompts while edited personas require fresh held-out generations. The held-out comparisons remain paired on identical contexts within each calibration–test seed pair.

Protocol freeze. Held-out metrics are reporting-only. No prompt, coordinate bound, search step, context seed, renderer wording, or candidate-selection rule is modified after inspecting the held-out results.

D.6 Cross-model and query-budget extension

Model cohort and interface gate. The cross-model extension freezes four local simulators: Gemma 3 4B, Gemma 3 12B, Qwen 3.5 9B, and Phi-4 14B. Before calibration, each model is evaluated on the same four deliberately clear persona–choice diagnostic pairs with three deterministic repeats per pair. All four models achieve parse success, repeat consistency, and directional accuracy of 1.0. Qwen 3.5 is run in its documented non-thinking chat-template mode; an initial diagnostic using its default reasoning mode exhausted the fixed response budget before emitting the requested JSON choice and is treated as an interface failure rather than a behavioral gate result.

Equal-budget cross-model calibration. All four models use calibration seed 1729, five context variants per intervention, coordinate step 0.5, bounds $[-1.5, 1.5]$, and the same coordinate enumeration. A full one-sweep search evaluates the exact authored population plus all 30 signed one-coordinate neighbors. Each population evaluation contains 120 persona–context queries, so the frozen full budget is 3,720 logical simulator calls per model. After a cohort-level freeze, each selected population is evaluated on four new held-out context seeds. Mean held-out error decreases for every model, but the selected move differs:

Screen–confirm query allocation. The fixed-order prefix diagnostic motivates a query-efficient allocation policy. Stage 1 evaluates the baseline plus all 30 signed coordinate moves on one of the five calibration-context variants, costing $31 \times 24 = 744$ logical simulator calls. Stage 2 fully evaluates the five best screened moves plus the baseline on the remaining four variants, costing $6 \times 96 = 576$ additional calls. The total budget is therefore 1,320 calls. An initially planned intermediate expansion to 2,496 calls is retained only as a diagnostic: it selects the same population as the 1,320-call policy for all four models, so the main method stops after selective confirmation.

The screen–confirm policy is designed only after the first cross-model held-out analysis and is therefore evaluated on a second fresh held-out seed set. It improves 15/16 model–context pairs with a 24.3% mean held-out gain on that diagnostic. To obtain a fair final strategy comparison, we then freeze five random-subset baselines and evaluate screen–confirm, fixed-prefix, and all random subsets on a third untouched held-out set.

Model	Selected persona	Feature	Step	Mean held-out gain
Gemma 3 4B	quality	brand	$+0.5$	22.7%
Gemma 3 12B	quality	brand	-0.5	14.8%
Qwen 3.5 9B	sustain	brand	-0.5	45.6%
Phi-4 14B	quality	brand	$+0.5$	12.0%

Table 21: **Frozen full-budget cross-model calibration.** The useful anchored correction is model-dependent; the table reports the full-budget selected coordinate and the mean reduction in weighted held-out RMSE on the first fresh cross-model evaluation set.

1,320-call strategy	Mean gain	Reliability	Mean held-out objective
Fixed prefix	-3.04%	6/16	0.1105
Random subset (5 seeds)	16.22%	80.0% avg.	0.0855
Screen–confirm	**20.87%**	**14/16**	**0.0805**

Table 22: **Final equal-budget strategy comparison on a third fresh held-out context set.** Random-subset values average five predeclared subsets of ten fully evaluated coordinate moves plus baseline. One random subset attains a larger mean gain than screen–confirm, so the result supports improved average efficiency and reliability rather than global optimality.

Adaptive outer search. After the third held-out comparison, we freeze two additional 1,320-call adaptive policies over the same 30 signed coordinate actions. Residual-prior LinUCB uses a fixed linear action embedding plus the economic residual direction as a weak prior and sequentially chooses ten fully evaluated moves. The history-aware LLM policy uses a fixed Gemma 3 12B proposer that sees the baseline residuals, the remaining action list, and all previously revealed calibration outcomes before selecting each next coordinate action. Both policies are frozen before a fourth fresh held-out cohort is generated.

Reflective evolutionary text search. After the fourth cohort is frozen and reported, we specify a final stress test with standalone GEPA 0.1.4. The original authored prompts remain immutable; GEPA may evolve only one suffix for each of the budget, quality, and sustainability personas. Reflective mutation receives variant-level economic residual guidance, while full candidate ranking uses the same five-variant weighted moment RMSE as the anchored coordinate experiments. Objective-level Pareto selection tracks budget-persona, quality-persona, sustainability-persona, and intervention-response blocks. A single deterministic Gemma 3 12B reflection model is used for all four behavioral simulators. The EIPG adapter independently counts logical behavioral simulator queries, and the frozen checkpoints choose the best candidate discovered by 1,320 and 3,720 calls.

Because GEPA changes the representation as well as the search algorithm, its comparison is deliberately interpreted as a stress test of whether richer reflective text evolution improves behavioral transfer. All GEPA candidates are frozen before the fifth held-out seeds are revealed.

Strategy	Calls	Mean held-out gain	Improved pairs	Mean objective
Fixed prefix	1 320	-6.9%	2/16	0.1055
History-aware LLM	1 320	-5.6%	1/16	0.1043
GEPA	1 320	-9.3%	3/16	0.1075
Random subset (5 seeds)	1 320	10.9%	73.8% avg.	0.0838
Residual-prior LinUCB	1 320	**19.6%**	12/16	**0.0756**
Screen–confirm	1 320	18.4%	**14/16**	0.0758
Full coordinate search	3 720	16.4%	13/16	0.0773
GEPA	3 720	5.7%	6/16	0.0906

Table 25: **Final outer-search stress test on a fifth untouched held-out cohort.** All candidates were frozen before these contexts were revealed. Random-subset statistics average five predeclared subsets. GEPA changes both optimizer and representation, so it is not an optimizer-only ablation. Lower objective is better.

Fourth-cohort strategy	Mean gain	Reliability	Mean objective
Fixed prefix	−8.30%	4/16	0.1050
History-aware LLM	−5.06%	3/16	0.1012
Random subset (5 seeds)	13.06%	78.8% avg.	0.0814
Residual-prior LinUCB	18.65%	12/16	0.0763
Screen–confirm	19.31%	15/16	0.0756

Table 23: **Adaptive-search comparison on the fourth fresh held-out cohort.** The history-aware LLM makes valid proposals but does not use the residual/history signal effectively; residual-prior LinUCB is competitive with screen–confirm at the same behavioral-query budget.

Model	GEPA 1,320	GEPA 3,720	Full coordinate 3,720
Gemma 3 4B	−36.8%	3.2%	19.5%
Gemma 3 12B	0.0%	0.0%	−4.8%
Qwen 3.5 9B	−0.4%	19.5%	43.2%
Phi-4 14B	0.0%	0.0%	7.6%
Cross-model mean	−9.3%	5.7%	16.4%
Improved pairs	3/16	6/16	13/16

Table 24: **GEPA text-space stress test on the fifth fresh held-out cohort.** Values are mean reductions in weighted held-out RMSE relative to the authored prompts. GEPA becomes useful at the larger budget for Qwen 3.5 9B and modestly for Gemma 3 4B, but transfer is less consistent than full anchored coordinate search.

Sequential freeze discipline. Five held-out cohorts serve distinct, sequential roles. The first evaluates the frozen full-budget cross-model cohort. After observing its fixed-prefix budget behavior, screen–confirm is specified and evaluated on a second fresh set. That result motivates five frozen random-subset baselines, which are compared with fixed-prefix and screen–confirm on a third fresh set. Residual-prior LinUCB and the history-aware LLM selector are then specified and evaluated on a fourth fresh set. Finally, GEPA is frozen and compared with all prior strategies on a fifth fresh set. No reported strategy is tuned on the held-out cohort used for its own final comparison; once the fifth-cohort GEPA outcomes are observed, outer-search experimentation stops.

E Public-Benchmark Respondent-Holdout Robustness

As a secondary robustness check, we use the predeclared respondent split in the frozen Nutri2Cycle configuration: 20% of eligible Belgium-bread respondents are held out using seed 27017, while Tasks 1–3 remain the only task outcomes used. Candidate search is performed on the remaining respondents; the held-out respondents are evaluated only after candidate selection. Task 4 is not read in this diagnostic.

Table 26: **Public Nutri2Cycle evaluation.** Panel A is the primary frozen Task-4 test: persona candidates are selected using Tasks 1–3 only and Task 4 is revealed afterward. Panel B is the predeclared 20% respondent holdout on Tasks 1–3. Lower is better.

Panel A: untouched Task 4

Method	Share L2 ↓	NLL ↓	Response L2 ↓
Initial static	0.6263	1.8705	0.0174
Anchor-only EIPG	0.4609 (0.0283)	1.5888	0.0348
Intervention-rich EIPG	**0.4491** (0.0229)	**1.5674**	**0.0181**
Direct human MNL	0.0758	1.1811	0.0696

Panel B: unseen respondents, Tasks 1–3

Method	Pooled share RMSE ↓	Response RMSE ↓
Initial static	0.3316	**0.0475**
Anchor-only EIPG	**0.2586**	0.0484
Intervention-rich EIPG	**0.2586**	0.0484

Both calibrated methods reduce pooled task-share RMSE from 0.3316 to 0.2586, a reduction of approximately 22.0% relative to the static generator. This shows that the calibration gain in choice levels is not confined to the respondents used for search. However, intervention-response RMSE is not improved in this diagnostic: the static generator scores 0.0475 versus 0.0484 for both calibrated methods. Moreover, anchor-only and intervention-rich calibration have identical aggregate held-out metrics here. We therefore treat this check as support for respondent-level robustness of the overall calibration effect, not as additional evidence that intervention-rich calibration dominates anchor-only calibration.

This appendix result is deliberately secondary to the primary public benchmark. The primary claim remains based on the frozen Task 4 evaluation, where intervention-rich EIPG modestly improves over anchor-only calibration on share L2 and NLL. The respondent-holdout diagnostic instead asks a different question: whether candidate selection on one subset of respondents transfers to unseen respondents under the same Tasks 1–3 design.