SAME CHOICE, DIFFERENT RESPONSE: EVALUATING LLM SIMULATORS UNDER ATTRIBUTE CHANGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used to simulate human choices, but reproducing observed choices may not predict how people respond when decision attributes change. We study *interventional response alignment*: agreement between human and model changes in the full choice-probability vector. We show that low observational choice risk does not, without additional assumptions, guarantee low response risk. On SwissMetro, we evaluate responses to travel-time and fare changes against fitted population-level human references. Prompted LLMs show response discrepancies even when their baseline choices match recorded human choices. More strikingly, on a fixed respondent-disjoint test set, choice-only SFT and DPO raise a 7B model's accuracy from 44.3% to 69.3–70.8%, yet improve time alignment while worsening fare alignment and the prespecified combined endpoint. Supervising paired response differences reduces the combined gap by 41% relative to SFT and outperforms choice augmentation on the same inputs. This gain does not extend to a held-out attribute, comes with worse negative log-likelihood, and has an unresolved accuracy difference. Because responses to the constructed changes are unobserved, these findings concern agreement with fitted population references rather than individual counterfactual fidelity. They show why choice accuracy alone is insufficient to evaluate simulators intended for use under change.

1 INTRODUCTION

Behavioral simulators are commonly evaluated by how well they reproduce human choices under observed conditions. Their intended use often requires a further step: predicting how choices change when prices, policies, service attributes, or available alternatives are modified. Discrete-choice models characterize such changes through attribute sensitivity and substitution patterns (McFadden, 1974; Ben-Akiva & Lerman, 1985; Train, 2009). Yet human responses are rarely observed for every proposed change. Controlled experiments are costly, and observational data cover only encountered contexts. A simulator can therefore fit recorded choices while responding unreliably to the changes it is used to evaluate (Lucas Jr, 1976).

Large language models (LLMs) offer a scalable approach to behavioral simulation. They have been used as synthetic survey respondents, experimental participants, and interactive agents (Argyle et al., 2023; Aher et al., 2023; Park et al., 2023; Dillion et al., 2023); specialized models also learn from trial-level choices or population response distributions (Binz et al., 2025; Cao et al., 2025). Evaluations primarily assess choice accuracy or distributional agreement in observed contexts. Such agreement can coexist with distorted subgroup relationships and response correlations, as well as sensitivity to question wording (Bisbee et al., 2024; Tjuatja et al., 2024; Williams et al., 2026). Whether improved choice fit also yields human-aligned responses to changes in decision attributes remains unclear.

The distinction matters because choices at observed contexts do not, in general, determine a simulator's behavior under an intervention. Two simulators can select the same baseline alternative, or assign the same baseline choice distribution, yet respond differently when travel time or fare changes. This is an instance of underspecification along dimensions relevant to deployment (D'Amour et al., 2022). Asking a model to explain its choice does not settle the question: generated explanations may

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.

not faithfully describe the factors driving its predictions (Turpin et al., 2023). We therefore ask: *Does learning to reproduce human choices also yield human-aligned responses to interventions?*

We study this question through *interventional response alignment*: agreement between human and model changes in the full choice-probability vector under specified attribute interventions. Comparing the full vector captures both sensitivity to the changed attribute and substitution among alternatives. We show formally that low observational choice risk does not, without additional assumptions, guarantee low response risk. We then test this distinction using SwissMetro, a stated-preference dataset of choices among Train, SwissMetro, and Car (Bierlaire et al., 2001). Its recorded attributes allow us to construct changes to travel time and fare. Human responses to these constructed changes are not directly observed, so we compare model responses with fitted population-level human references. We assess those references using respondent-disjoint prediction, alternative specifications, and near-matched within-respondent comparisons. Our empirical claims concern agreement with these fitted references, not individual causal counterfactuals or internal cognitive mechanisms.

Our experiments yield three findings. First, prompted LLMs show response discrepancies even when they reproduce the recorded human choice. The discrepancies persist across reference specifications, intervention magnitudes, semantic reformulations, and model families. Second, improved choice prediction does not consistently improve response alignment. On a fixed respondent-disjoint test cohort, supervised fine-tuning (SFT) and direct preference optimization (DPO) raise Qwen2.5-7B choice accuracy from 44.3% to 69.3–70.8%, while improving time alignment but worsening fare alignment and the prespecified combined time–fare endpoint. This attribute-specific pattern persists across seeds, training-exposure controls, DPO settings, and calibration analyses. Third, supervising paired intervention-induced probability changes reduces the mean combined gap from 0.0358 under SFT to 0.0212, a 41% reduction, and outperforms choice augmentation using the same intervention inputs. The gain has limits: the accuracy difference is statistically unresolved, negative log-likelihood worsens, and response alignment does not improve on a held-out attribute. The training objective is related to sensitivity matching in Sobolev training (Czarnecki et al., 2017), but targets finite probability differences rather than analytic derivatives.

As complementary evidence, we examine matched shopping actions in OPeRA. Human and model consideration sets agree exactly in 12.9% of 225 action matches, with a small positive Jaccard lift over an action- and length-compatible null. This exploratory result concerns expressed rationales; it is not a replication of the intervention analysis.

Our contributions are:

- We define interventional response alignment using changes in the full choice-probability vector and establish why observational choice fit alone cannot guarantee it.
- We develop an evaluation protocol that combines fitted human-response references, baseline-matched diagnosis, fixed-cohort training comparisons, and checks of reference and elicitation sensitivity.
- We show that choice learning can improve accuracy while worsening attribute-specific response alignment, and that direct response supervision can reduce the targeted gap beyond choice augmentation on the same inputs.

2 Related Work

LLMs as human simulators. LLMs have been used as synthetic survey respondents, experimental participants, and interactive agents (Argyle et al., 2023; Aher et al., 2023; Park et al., 2023). Prior work evaluates population variation (Bisbee et al., 2024), persona-conditioned prediction (Hu & Collier, 2024; Lutz et al., 2025), interactive user behavior (Yoon et al., 2024), and behavioral signatures across tasks (Coda-Forno et al., 2024; Binz et al., 2025). These studies primarily evaluate behavior at observed contexts, while our work examines responses to attribute changes.

Beyond observational fit. Similar predictive performance can conceal deployment-relevant differences (D'Amour et al., 2022). LLMs may diverge from humans under survey reformulations (Tjuatja et al., 2024), and accurate marginal distributions may coexist with inaccurate response correlations (Williams et al., 2026). Related work studies behavioral testing (Ribeiro et al., 2020), statistical

calibration of synthetic participants (Hullman et al., 2026), and intervention-induced user drift (Lin et al., 2026). We evaluate human–model response agreement using discrete-choice references for attribute sensitivity and substitution (McFadden, 1974; Train, 2009).

Learning response functions. Existing methods learn human choices or response distributions at observed contexts (Binz et al., 2025; Cao et al., 2025), including through preference optimization (Rafailov et al., 2023). Our response-aware objective instead supervises intervention-induced finite differences and is related to Jacobian matching (Srinivas & Fleuret, 2018). This allows us to test whether response supervision adds value beyond choice learning.

3 Background and Problem Setup

We study a behavioral simulator used to predict how human choices respond to changes in a decision environment. The available data record choices under observed conditions, while the intended application concerns choices under modified conditions. This section defines the choice distributions, their responses to specified changes, and the human response reference used in our empirical evaluation. Section 4 then defines response disagreement and the corresponding evaluation protocol.

3.1 Decision Setting

A decision context x contains information about a person and the alternatives presented to them. In travel choice, it may include traveler characteristics and the time, fare, and headway of each mode. Let $\mathcal{A}$ be the finite set of possible alternatives, with availability specified by x. The observational dataset is

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}, \qquad y_i \in \mathcal{A}, \tag{1}$$

where multiple observations may come from the same respondent.

For $j \in \{H, M\}$, let

$$p_j(x) = \big(p_j(a \mid x)\big)_{a \in \mathcal{A}} \tag{2}$$

denote the human (H) or model (M) choice distribution at x. Its entries sum to one, and unavailable alternatives have probability zero. We use *policy* to mean this mapping from a represented context to a choice distribution. The term does not imply a sequential decision process or a complete account of human cognition.

A recorded choice y_i is one realization from the human choice distribution. It need not equal the distribution's modal choice. For a probabilistic policy, define

$$f_j(x) = \arg\max_{a \in \mathcal{A}} p_j(a \mid x), \tag{3}$$

with a fixed tie-breaking rule. Two policies can have the same modal choice without assigning the same probabilities to the alternatives. Even equality of their full distributions at one context does not determine their distributions at a modified context.

3.2 Attribute Interventions and Responses

An intervention $I_k \in \mathcal{I}$ maps a baseline context x to a modified context $I_k(x)$. The index k specifies the attribute, alternative, and magnitude of the change. For example, a fare intervention changes one mode's fare while keeping the other represented attributes fixed. In this paper, an intervention is a controlled change to the simulator's input. It does not, on its own, identify a real-world causal effect or establish that unobserved determinants of human choice remain fixed.

The response of policy j to I_k is the change in its choice-probability vector:

$$\Delta p_k^j(x) = p_j(I_k(x)) - p_j(x), \qquad j \in \{H, M\}. \tag{4}$$

The components of this vector sum to zero. It captures the direction and magnitude of the change for each alternative, including substitution among alternatives. Two policies can retain the same modal choice after an intervention while producing different response vectors.

We call agreement between $\Delta p_k^M(x)$ and $\Delta p_k^H(x)$ over specified interventions *interventional response alignment*. The target concerns observable choice distributions under change; it does not imply agreement in internal reasoning or cognitive mechanisms.

3.3 HUMAN RESPONSE REFERENCE

The ideal comparison uses the human conditional response $\Delta p_k^H(x)$. The observational data, however, generally do not contain human choices under our constructed interventions. In particular, the recorded baseline choice y_i cannot serve as the outcome label at $I_k(x_i)$.

We estimate a population-level human choice function $\widehat{p}_H$ and construct the empirical reference

$$\widehat{\Delta p}_k^{\,H}(x) = \widehat{p}_H(I_k(x)) - \widehat{p}_H(x). \tag{5}$$

Our empirical comparison thus measures agreement with a fitted human response, subject to the adequacy of the fitted reference and support for the modified contexts. It does not establish an individual's unobserved counterfactual choice. In the next section, we define the response gap relative to this reference, distinguish baseline-matched diagnosis from fixed-cohort training comparisons, and specify how reference uncertainty enters the evaluation.

4 METHODOLOGY

Section 3 defined the human and model response vectors under a specified intervention. We now measure their disagreement, examine what observational choice fit can establish, and define the population endpoint used in our experiments.

4.1 MEASURING RESPONSE ALIGNMENT

For a context x and intervention I_k, define the context-level response gap

$$G_k(x) = \frac{1}{2}\left\|\Delta p_k^M(x) - \Delta p_k^H(x)\right\|_1. \tag{6}$$

A smaller gap indicates closer agreement in response direction, magnitude, and substitution across alternatives. Because the quantities inside the norm are *differences* of probability vectors, $G_k(x)$ can lie in $[0, 2]$; the factor $1/2$ does not make it a total-variation distance between two probability distributions.

We use *interventional response alignment* to describe small response gaps over a specified set of contexts and interventions. Alignment is therefore relative to the attributes, intervention magnitudes, and population being evaluated. Stability under meaning-preserving prompt changes is assessed separately as an elicitation robustness check. Neither a small gap nor stability under paraphrase identifies the model's internal reasoning or a human cognitive mechanism.

4.2 WHAT CHOICE FIT CAN ESTABLISH

Proposition 1 (single-context non-identification). Suppose $I_k(x)$ is unconstrained by observations at x. Agreement in the modal baseline choice, $f_H(x) = f_M(x)$, does not bound $G_k(x)$. Even if $p_H(x) = p_M(x)$, it is possible to have $G_k(x) = 1$.

See Appendix G.1 for the proof.

This example concerns one context. The same issue can persist with many observed contexts when proposed interventions reach contexts whose responses are not constrained by the observational data.

Proposition 2 (choice risk does not determine response risk). Let P_{XY} be the observational distribution of contexts and choices, and let P_I be a distribution over baseline context–intervention pairs. Define

$$R_{\mathrm{ch}}(p) = \mathbb{E}_{(x,y)\sim P_{XY}}\,\ell\big(p(x), y\big), \tag{7}$$

$$R_{\mathrm{rsp}}(p) = \mathbb{E}_{(x,I)\sim P_I}\frac{1}{2}\left\|[p(I(x)) - p(x)] - [p_H(I(x)) - p_H(x)]\right\|_1. \tag{8}$$

If interventions can reach contexts outside the support constrained by P_{XY}, observational choice risk alone provides no general guarantee that response risk approaches zero. In particular, for unrestricted policies and a strictly proper choice loss, there are settings with a sequence p_n such that

$$R_{\mathrm{ch}}(p_n) - R^\star_{\mathrm{ch}} > 0, \qquad R_{\mathrm{ch}}(p_n) - R^\star_{\mathrm{ch}} \longrightarrow 0,$$

while $R_{\mathrm{rsp}}(p_n)$ remains bounded away from zero.

See Appendix G.2 for the proof.

The proposition describes a limit of the observational objective. It does not claim that choice-trained models cannot learn useful responses through inductive bias, or that lack of observational support explains every gap measured below.

Proposition 3 (endpoint-error bound). Suppose the model's distributions at the baseline and intervened contexts satisfy

$$\|p_M(x) - p_H(x)\|_1 \le \epsilon_0, \qquad \|p_M(I_k(x)) - p_H(I_k(x))\|_1 \le \epsilon_k.$$

Then

$$G_k(x) \le \frac{\epsilon_0 + \epsilon_k}{2}. \tag{9}$$

See Appendix G.3 for the proof.

A large gap against the *true* human policy would rule out accurate simulation at both endpoints of that context–intervention pair. A small gap is a weaker result: baseline and intervened errors may offset one another, and agreement on tested responses does not establish agreement elsewhere or reveal the mechanism producing them. Empirically, we use a fitted human reference, so interpretation of a measured gap also depends on the adequacy of that reference.

4.3 Empirical Endpoint and Evaluation Cohorts

The propositions concern a response at one context. Our primary empirical endpoint compares *population-average response vectors* on a specified cohort $\mathcal{C}$. For fixed weights $w_i \ge 0$ with $\sum_{i\in\mathcal{C}} w_i = 1$, it is

$$\widehat{G}_{k,\mathcal{C}} = \frac{1}{2}\left\| \sum_{i\in\mathcal{C}} w_i \left[\Delta p_k^M(x_i) - \widehat{\Delta p}_k^{\,H}(x_i)\right] \right\|_1. \tag{10}$$

The fitted reference $\widehat{\Delta p}_k^{\,H}$ is defined in Eq. (5). With an ideal human reference, the triangle inequality gives

$$G_{k,\mathcal{C}} \le \sum_{i\in\mathcal{C}} w_i G_k(x_i).$$

Thus a small aggregate gap can conceal errors that offset across contexts. Equation (10) measures population response agreement on the specified cohort; it does not establish accurate responses for each person or scenario.

We use two cohort rules for different questions. The *same-choice diagnostic* conditions on cases where the model's baseline action matches the recorded human choice:

$$\mathcal{C}_M = \{i : \widehat{y}_i^M = y_i\}. \tag{11}$$

Here $\widehat{y}_i^M$ is the model's modal action when probabilities are available, or its elicited action in a discrete-response experiment. Matching y_i does not imply that the model matches $p_H(x_i)$, or even the human distribution's modal choice. Because this rule can select an unrepresentative subset, we report matched-cohort composition and assess selection sensitivity.

For *training comparisons*, every policy is evaluated on the same full held-out cohort. This prevents a change in baseline accuracy from changing the cases included in its response endpoint. Trained models contribute answer probabilities to Eq. (10). For prompted models that provide only elicited choices, one-hot actions are aggregated into choice shares; their individual outputs are not treated as estimates of complete latent choice distributions.

4.4 Response-Supervised Training

Proposition 2 motivates testing whether supervision on responses provides information that choice labels alone leave unconstrained. For training contexts x_i, let

$$r_i^k = \widehat{p}_H(I_k(x_i)) - \widehat{p}_H(x_i)$$

be a response target produced by a human reference fitted using training respondents only. We augment baseline choice loss with

$$\mathcal{L}_{\mathrm{RA}} = \mathcal{L}_{\mathrm{ch}} + \lambda\, \mathbb{E}_{i,k} \frac{1}{2} \left\| [p_\theta(I_k(x_i)) - p_\theta(x_i)] - r_i^k \right\|_1 . \tag{12}$$

The choice term supervises behavior at the observed baseline; the second term directly supervises finite differences between baseline and intervened probabilities. Its targets are fitted human responses, so improvement would show that this additional supervision helps on the tested reference and attributes. It would not identify a unique cause of the original gap or guarantee transfer to other interventions.

Merely presenting intervention inputs during training does not impose the finite-difference constraint in Eq. (12): policies can share hard choice labels while assigning different response vectors. We therefore compare response-aware training with choice augmentation on the same inputs. We also test whether validation-fitted scalar temperature scaling accounts for response gaps associated with probability calibration. These controls isolate specific alternatives to the response-supervision explanation.

5 Experimental Design

Our experiments address three questions: whether response gaps remain beneath matched baseline choices, whether learning observed choices improves response alignment, and whether response supervision helps beyond exposure to intervention inputs. SwissMetro provides the primary intervention tests. OPeRA supplies a separate, exploratory analysis of expressed considerations under matched actions.

5.1 SwissMetro and Human Response References

SwissMetro contains 10,719 choices from 1,191 respondents among Train, SwissMetro, and Car. We modify SwissMetro time and fare while holding other recorded attributes fixed; luggage serves as a secondary, held-out transfer attribute.

We evaluate responses against fitted population references because human choices under these changes are unobserved. On respondent-disjoint folds, the primary and nonlinear references must outperform pooled multinomial logit in held-out log loss and recover the expected time and fare directions. Near-matched within-respondent choices provide an additional check. For training comparisons, the reference is fitted using training respondents only.

5.2 Response Gaps Beneath Matched Choices

We compare prompted LLMs with fitted human references on cases where their baseline choice matches the recorded choice (Eq. (11)), using aggregate response gaps (Eq. (10)). We report cohort composition because matching is selective.

Exact repeats and paraphrases test elicitation stability. A fixed 824-case sample tests time and fare changes from -30% to $+30\%$. We also compare two human references and independently served Qwen and DeepSeek models on frozen scenarios, including a common-matched cohort.

5.3 Transfer From Choice-Only Learning

We next test whether learning recorded choices improves response alignment. Respondents are assigned to disjoint training, validation, and test sets containing 7,506, 1,602, and 1,611 choices, respectively. Starting from Qwen2.5-7B-Instruct, we train supervised fine-tuning (SFT) and direct preference optimization (DPO) policies using baseline choices only. Neither objective receives intervention examples. We evaluate the base model and every adapted policy on the same full test cohort, avoiding changes in cohort composition as accuracy improves.

Choice fit is measured by accuracy and negative log-likelihood. The prespecified response endpoint is the mean of the time and fare population gaps,

$$G_{\text{core}} = \frac{\widehat{G}_{\text{time},\mathcal{C}} + \widehat{G}_{\text{fare},\mathcal{C}}}{2}, \tag{13}$$

where $\mathcal{C}$ is the fixed test cohort. Attribute-specific gaps remain visible alongside this combined endpoint. We repeat the primary SFT and DPO conditions across three seeds. A near-step-matched SFT run checks training exposure, DPO coefficient variants check setting sensitivity, and validation-fitted temperature scaling tests whether probability calibration accounts for the response pattern.

5.4 Response-Supervised Repair

Finally, we train with the response-aware objective in Eq. (12), using baseline choice labels and fitted time and fare response targets from training respondents. The response-loss weight is selected on validation data under a prespecified accuracy constraint; test data are used only after selection. Luggage is excluded from response supervision and evaluated for cross-attribute transfer.

To distinguish response targets from intervention-input exposure, a choice-augmentation control sees the same baseline, time, and fare training contexts but receives hard choice supervision rather than response-vector targets. We compare these policies on the same frozen test cohort, reporting choice fit, the combined response gap, and each attribute gap. Model contrasts use paired respondent-clustered bootstrap resampling; a separate analysis refits the human reference to assess reference-estimation uncertainty.

5.5 Supporting Analysis of Expressed Considerations

OPeRA tests a related but distinct question. We compare coded human and model consideration sets only where their shopping actions match, using action- and rationale-length-compatible comparisons. This analysis concerns expressed rationales. It does not provide human outcomes under attribute interventions and is not treated as a replication of the SwissMetro response tests.

6 Results

6.1 Human Response References

Both fitted references improve respondent-disjoint log loss over pooled MNL and recover the expected time and fare response directions in all five folds (Table 26). Near-matched within-respondent checks retain 596 pairs per attribute. Across 256 observed switches, direction agreement is 64.5% for the primary reference and 66.0% for the nonlinear reference, with corresponding rank correlations of .20 and .22. These checks support their use as population response benchmarks, while the moderate agreement and possible errors shared by both specifications limit stronger interpretations. The results below measure alignment with fitted human responses, not directly observed responses to the constructed interventions.

Table 1: Human-reference validation on respondent-disjoint folds.

Reference	Accuracy ↑	Log loss ↓	Time/fare signs
Pooled MNL	.658	.810	5/5
Primary	.650	.769	5/5
Nonlinear	.655	.756	5/5

Arrows indicate the preferred direction; signs are checked in all five folds.

6.2 Response Discrepancies Beneath Matched Choices

Response discrepancies remain when a prompted model reproduces the recorded baseline choice (Table 2). Against the nonlinear reference, Moonshot's SwissMetro-share response is smaller by 6.49 percentage points for time and 4.90 points for fare. Gaps persist under paraphrasing and across

intervention magnitudes. Fare changes also induce more switching than exact repetition on the same 824 cases. On 200 common baseline matches, independently served Qwen and DeepSeek models both show fare gaps beyond their repeat-placebo gaps.

Table 2: Response discrepancies beneath matched choices.

Comparison / cohort	Result
Moonshot baseline matches	Signed error: time -6.49 pp [-6.96, -6.03]; fare -4.90 pp [-5.57, -4.24].
Paraphrase; 679/824 rematches	Time/fare gaps: $G = .050/.042$.
Repeat vs. fare; fixed 824	Switching: 1.0% vs. 5.1%; paired excess: 4.0 pp [2.5, 5.7].
Dose curves; fixed 824	Mean G: time .048 [.043, .053]; fare .031 [.027, .038].
Qwen / DeepSeek; 200 common matches	Fare gap minus repeat: .227 [.149, .321] / .139 [.088, .217].

G: half-L_1 response gap; pp: percentage points. Brackets give 95% respondent-clustered intervals. Rows use distinct cohorts. Signed errors compare SwissMetro-share responses with the nonlinear reference.

The dose curves show that Moonshot's response remains flatter across changes from -30% to $+30\%$: its time and fare slopes are $-.049$ and $-.090$, compared with $-.267$ and $-.193$ for the primary reference. Earlier and contemporaneous anchors agree on at least 98.2% of weighted choices, reducing concern about batch instability. Under-response is not a common direction across models, however; Kimi can over-react where Moonshot under-reacts.

Baseline matching is selective. Moonshot matches 38.3% of choices overall but only 3.6% of Swiss-Metro choices. Reference refitting and common-cohort reweighting retain response discrepancies, but the matched estimates remain conditional on model success. The next experiment uses a fixed full test cohort to assess what happens as choice accuracy improves.

6.3 Choice Learning and Response Alignment

Choice-only training raises test accuracy from .443 for the base model to approximately .701 for SFT and DPO (Table 3). Over the same fixed cohort, both methods improve time alignment and worsen fare alignment. Their mean core gaps increase from .030 for Base to .036 and .049, respectively. Thus the combined endpoint reflects opposing attribute-level changes, with the fare deterioration outweighing the time improvement.

Table 3: Choice fit and response alignment on the frozen test cohort.

	Choice fit		Response gap ↓		
Training	Accuracy ↑	NLL ↓	Time	Fare	Core
Base	.443	1.330	.025	.034	.030
SFT	.701±.005	.669±.004	.016	.056	.036±.003
DPO	.701±.008	3.270±.104	.004	.095	.049±.007
Choice augmentation†	.688	.705	.015	.047	.031
Response-aware	.690±.005	.699±.001	.016	.026	.021±.001

Three-seed entries are mean±SD; † denotes the seed-22 exposure ablation. Time and fare entries for three-seed conditions are rounded means. Full intervals appear in Tables 6 and 7.

The separation persists under controls for training exposure and calibration. Near-step-matched SFT reaches .709 accuracy but a .042 core gap. Validation-fitted temperature scaling lowers the primary-β DPO runs' NLL to .665–.683, while their core gaps remain .036–.038, compared with .029 for calibrated Base. The calibration result has a boundary: at $\beta = .3$, the calibrated core increase is unresolved (.0046 [$-.0010$, .0103]), although the fare gap still increases by .0438 [.0338, .0536]. The evidence therefore supports an attribute-specific time–fare asymmetry across these settings, without

implying that every calibrated condition significantly worsens the combined endpoint (Tables 13 and 22).

An exploratory GA-pass analysis gives a concrete example of the fare discrepancy. DPO responds to a lower displayed fare even for pass holders whose effective fare remains zero (Table 23). Nonholders also show fare gaps, so this behavior does not account for the entire result.

6.4 Response-Supervised Training

Validation selects the response-loss weight $\lambda = 1$; $\lambda = 5$ misses the prespecified accuracy floor. Across three seeds, response-aware training reduces the mean core gap from .0358 under SFT to .0212, a 41% reduction (Table 4). The gain is driven primarily by fare. On the seed-22 same-input comparison, choice augmentation attains a .0308 core gap, versus .0224 with response supervision. Because these conditions use the same intervention inputs, steps, and forward passes, input exposure alone does not explain their difference.

Table 4: Response supervision: paired gains and limits.

Comparison	Endpoint	Difference [95% CI]
RA minus SFT	Core gap	$-.0146$ [$-.0168$, $-.0125$]
	Accuracy	$-.0106$ [$-.0275$, .0066]
	NLL	.0302 [.0087, .0521]
Augmentation minus RA	Core gap	.0084 [.0076, .0093]

RA denotes response-aware training. RA–SFT contrasts average seed-matched comparisons across three seeds; the augmentation comparison uses seed 22. Intervals are paired and respondent-clustered.

This targeted gain has costs and limited transfer. The accuracy difference from SFT is unresolved, NLL worsens, and alignment on held-out luggage does not improve. The experiment establishes an advantage for supervising the tested time and fare responses against the fitted reference; it does not establish a generally more faithful choice policy.

6.5 Expressed Considerations Under Matched Actions

In OPeRA, 225 of 604 shopping actions match between the human and model. Among these matches, consideration sets agree exactly in 12.9% of cases; mean Jaccard overlap is .424 versus .390 under an action- and length-compatible null ($p = .0007$). Matched actions therefore carry some information about expressed considerations, although identical consideration sets are uncommon. This single-coder analysis concerns rationales and does not independently test intervention responses. Additional results appear in Appendix D.

7 Discussion

Improved choice accuracy does not guarantee aligned responses to attribute changes. Choice-only training improves time alignment but worsens fare alignment, while direct response supervision reduces the targeted gap beyond training on the same intervention inputs. The gain does not transfer to held-out luggage and comes with worse NLL. Simulators used for pricing or service evaluation should therefore be tested on the specific responses those applications require.

These findings concern population-average agreement with fitted human references, since human choices under the constructed changes are unobserved. Aggregate errors may offset, and baseline matching selects a limited cohort. Training covers one 7B model family and one decision domain; the same-input ablation uses one seed. OPeRA provides complementary evidence on expressed rationales, not intervention responses. Validation against directly observed human responses across more settings remains necessary.

AI USE STATEMENT

Generative AI tools were used to assist with literature discovery, research brainstorming, code drafting and debugging, statistical-analysis scripts, figure and table preparation, and drafting and editing portions of the paper. The authors inspected the source literature, executed and checked the analysis code, verified reported quantities against saved artifacts, and reviewed all AI-assisted text and outputs. The empirical claims are based on the documented data, models, and analyses rather than unverified model-generated assertions. The authors take responsibility for the final content, including all text, claims, code, and artifacts produced with AI assistance.

ETHICS STATEMENT

This work evaluates the risks of treating language models as human simulators. The motivating sports dataset was collected directly by the researchers from 24 volunteers. Before participating, they were explicitly informed that their responses would be used for academic research and voluntarily agreed to take part; participation was uncompensated. The analysis uses their decision responses and accompanying rationales in aggregate form and does not seek to diagnose individual participants. SwissMetro and OPeRA are analyzed under their documented access conditions; the released analysis should exclude direct personal identifiers. The fitted human reference is a population-level statistical model rather than an individual causal or psychological diagnosis. Results should therefore not be used to infer the cognition of particular people or to justify consequential transportation, pricing, or product decisions without domain-specific validation. The institutional-review or exemption status of the sports collection must be stated precisely in the submission record.

REPRODUCIBILITY STATEMENT

The appendix documents the respondent-disjoint split, human-reference models, intervention construction, prompts, matching estimands, training objectives, hyperparameters, calibration controls, and respondent-clustered bootstrap. The supplementary archive contains immutable split identifiers, row-level scores, analysis outputs, and scripts that regenerate the reported tables and figures. Model checkpoints and third-party data are distributed only where their licenses and access conditions permit.

REFERENCES

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

Moshe E Ben-Akiva and Steven R Lerman. *Discrete choice analysis: theory and application to travel demand*, volume 9. MIT press, 1985.

Michel Bierlaire, Kay W. Axhausen, and Georg Abay. The acceptance of modal innovation: The case of Swissmetro. In *Proceedings of the 1st Swiss Transport Research Conference*, Monte Verità, Ascona, Switzerland, 2001. URL `https://transp-or.epfl.ch/documents/proceedings/BierAxhaAbay01.pdf`.

Marcel Binz, Zeynep Akata, Matthias Bethge, and Eric Schulz. A foundation model to predict and capture human cognition. *Nature*, 2025.

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

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations.

In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2025.

Julian Coda-Forno, Marcel Binz, Jane X. Wang, and Eric Schulz. CogBench: A large language model walks into a psychology lab. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 9076–9108, 2024.

Wojciech M. Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/758a06618c69880a6cee5314ee42d52f-Abstract.html.

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in cognitive sciences*, 27(7):597–600, 2023.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in llm simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 10289–10307, 2024. doi: 10.18653/v1/2024.acl-long.554.

Jessica Hullman, David Broska, Huaman Sun, and Aaron Shaw. This human study did not involve human subjects: Validating LLM simulations as behavioral evidence. *arXiv preprint arXiv:2602.15785*, 2026. Preprint.

Victoria Lin, Taedong Yun, Maja Matarić, John Canny, Arthur Gretton, and Alexander D'Amour. The illusion of intervention: Your LLM-simulated experiment is an observational study. *arXiv preprint arXiv:2605.20767*, 2026. Preprint.

Robert E Lucas Jr. Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, volume 1, pp. 19–46. North-Holland, 1976.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23212–23237, 2025. doi: 10.18653/v1/2025.findings-emnlp.1261.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka (ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press, 1974.

Joon Sung Park, Joseph 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*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020.

Suraj Srinivas and Francois Fleuret. Knowledge transfer with jacobian matching. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4723–4731, 2018.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024. doi: 10.1162/tacl_a_00685.

Kenneth E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2 edition, 2009.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf.

Tristan Williams, Franziska Weeber, Sebastian Padó, and Alan Akbik. Beyond marginal distributions: A framework to evaluate the representativeness of demographic-aligned LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4806–4826, 2026. doi: 10.18653/v1/2026.findings-acl.236.

Se-eun Yoon, Zhankui He, Jessica Echterhoff, and Julian McAuley. Evaluating large language models as generative user simulators for conversational recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 1490–1504, 2024. doi: 10.18653/v1/2024.naacl-long.83.

A Appendix Guide

The appendix contains the response-aware and robustness details, a compact sports preliminary, expanded SwissMetro and OPeRA analyses, and complete preference-learning configuration. Earlier sports-draft chapters and the superseded heuristic SwissMetro simulation are preserved in the source archive but excluded from the submission build. The compact replacement consistently treats sports labels as expressed considerations rather than latent cognition.

A.1 Frozen response-aware learning extension

Before scoring any response-aware checkpoint on validation or test data, we froze an extension that optimizes both the observed baseline choice and the intervention response. For training scenario i, let p_θ^0 be the model distribution at baseline, let p_θ^k be its distribution after $k \in \{\text{time}, \text{fare}\}$, and let r_i^k be the corresponding response of the training-only human reference. The objective is

$$\mathcal{L}_{\text{RA}} = \text{CE}(p_\theta^0, y_i) + \lambda \frac{1}{2} \sum_k \left\| (p_\theta^k - p_\theta^0) - r_i^k \right\|_1 / 3. \tag{14}$$

The initial candidate fixes $\lambda = 5$, seed 22, one epoch, LoRA rank 16, and learning rate 3×10^{-6}. The human targets are produced only from training respondents. Model selection minimizes validation time–fare core gap subject to accuracy lying within two percentage points of choice-only SFT; only the selected setting is evaluated on the fixed test. Luggage is withheld from the response loss and serves as a cross-attribute transfer diagnostic. The dated freeze record also documents a validation-free numerical-stability amendment following an aborted FP16/MPS pilot.

Table 5: Validation-only response-aware result for the initial $\lambda = 5$ candidate. Response vectors follow (Train, SwissMetro, Car). Time and fare enter the training loss; luggage is held out as a transfer diagnostic. The candidate is excluded from confirmatory test scoring because accuracy .6979 falls below the frozen .7028 floor.

Quantity	Train	SwissMetro	Car
Human time response	−.0142	.0470	−.0328
Model time response	−.0051	.0410	−.0359
Human fare response	−.0060	.0317	−.0258
Model fare response	−.0068	.0354	−.0286
Human luggage response	−.0298	.0267	.0030
Model luggage response	−.0032	.0042	−.0010
Accuracy .6979; NLL .6483; Brier .4047			
Gaps: time .0091; fare .0037; core .0064; luggage .0266			

Table 6: Three-seed confirmatory and replication results. The $\lambda = 1$ setting was selected before test scoring; seeds 11 and 33 repeat the fixed protocol. Intervals use 10,000 respondent-clustered draws. Averaged over seed-matched contrasts, core gap changes by −.0146 [−.0168,−.0125] and accuracy by −.0106 [−.0275,.0066].

Condition	Accuracy	NLL	Brier	Core gap	Fare gap
base	0.443 [0.403, 0.482]	1.330 [1.230, 1.438]	0.759 [0.706, 0.814]	0.030 [0.026, 0.033]	0.034 [0.028, 0.039]
response_lambda1_seed11	0.685 [0.646, 0.722]	0.699 [0.638, 0.767]	0.425 [0.386, 0.467]	0.021 [0.018, 0.023]	0.023 [0.019, 0.028]
response_lambda1_seed22	0.695 [0.656, 0.731]	0.699 [0.631, 0.774]	0.422 [0.381, 0.467]	0.022 [0.019, 0.025]	0.031 [0.026, 0.036]
response_lambda1_seed33	0.691 [0.652, 0.729]	0.698 [0.634, 0.768]	0.424 [0.385, 0.467]	0.020 [0.018, 0.023]	0.024 [0.019, 0.028]
sft_seed11	0.701 [0.664, 0.737]	0.668 [0.604, 0.739]	0.403 [0.364, 0.445]	0.033 [0.030, 0.037]	0.051 [0.045, 0.058]
sft_seed22	0.705 [0.667, 0.742]	0.672 [0.601, 0.752]	0.402 [0.361, 0.447]	0.038 [0.034, 0.043]	0.061 [0.054, 0.069]
sft_seed33	0.696 [0.658, 0.732]	0.665 [0.601, 0.736]	0.400 [0.361, 0.441]	0.035 [0.032, 0.039]	0.054 [0.048, 0.061]

B Reproducibility Snapshot

All condition labels, seeds, prompts, split identifiers, checkpoints, and scored row-level probabilities are retained with the analysis archive. Every frozen robustness-matrix condition completed; failed or interrupted runs were resumed from checkpoints rather than silently replaced.

Table 7: Intervention-exposure ablation on the frozen test. Choice augmentation uses the same baseline, time, and fare training contexts as response-aware training, but applies only hard-choice cross-entropy. The seed-22 paired choice-augmentation-minus-response-aware differences are .0084 [.0076,.0093] for core gap and .0161 [.0147,.0175] for fare gap; the accuracy difference is $-.0062$ [$-.0199$,.0074]. Intervals use 10,000 respondent-clustered draws. Lower NLL and gaps are better.

Condition	Accuracy	NLL	Brier	Core gap	Fare gap
base	0.443 [0.403, 0.482]	1.330 [1.230, 1.438]	0.759 [0.706, 0.814]	0.030 [0.026, 0.033]	0.034 [0.028, 0.039]
choice_aug_seed22	0.688 [0.650, 0.726]	0.705 [0.638, 0.780]	0.426 [0.385, 0.470]	0.031 [0.027, 0.034]	0.047 [0.041, 0.053]
response_lambda1_seed11	0.685 [0.646, 0.722]	0.699 [0.638, 0.767]	0.425 [0.386, 0.467]	0.021 [0.018, 0.023]	0.023 [0.019, 0.028]
response_lambda1_seed22	0.695 [0.656, 0.731]	0.699 [0.631, 0.774]	0.422 [0.381, 0.467]	0.022 [0.019, 0.025]	0.031 [0.026, 0.036]
response_lambda1_seed33	0.691 [0.652, 0.729]	0.698 [0.634, 0.768]	0.424 [0.385, 0.467]	0.020 [0.018, 0.023]	0.024 [0.019, 0.028]
sft_seed11	0.701 [0.664, 0.737]	0.668 [0.604, 0.739]	0.403 [0.364, 0.445]	0.033 [0.030, 0.037]	0.051 [0.045, 0.058]
sft_seed22	0.705 [0.667, 0.742]	0.672 [0.601, 0.752]	0.402 [0.361, 0.447]	0.038 [0.034, 0.043]	0.061 [0.054, 0.069]
sft_seed33	0.696 [0.658, 0.732]	0.665 [0.601, 0.736]	0.400 [0.361, 0.441]	0.035 [0.032, 0.039]	0.054 [0.048, 0.061]

Table 8: Paired exact-repeat placebo and substantive interventions on the fixed 824-case dose-response sample. Entries are percentage points with 95% respondent-clustered bootstrap intervals. The normalized L_1 column measures the magnitude of the aggregate choice-share response.

Condition	Switch rate	$\Delta p(\mathrm{SM})$	Normalized L_1 magnitude
Exact-repeat placebo	1.0 [0.5, 1.7]	-0.1 [-0.5, 0.3]	0.2 [0.1, 0.8]
Time -20%	2.6 [1.5, 3.9]	0.2 [-0.2, 0.7]	0.5 [0.1, 1.7]
Fare -20%	5.1 [3.5, 6.8]	1.3 [0.4, 2.3]	1.4 [0.8, 2.9]
Time -20% minus placebo	1.6 [0.5, 2.8]	0.3 [-0.3, 0.9]	0.3 [-0.4, 1.4]
Fare -20% minus placebo	4.0 [2.5, 5.7]	1.4 [0.4, 2.5]	1.2 [0.4, 2.6]

Table 9: Selection robustness on the Qwen–DeepSeek common-matched cohort. Entries are intervention gap minus exact-repeat gap under the primary human reference, with respondent-clustered 95% intervals. Equal weighting assigns one third of mass to each baseline choice before applying within-stratum sample weights.

Weighting/cohort	Qwen time	Qwen fare	DeepSeek time	DeepSeek fare
Equal baseline-choice weight	.068 [.026,.119]	.192 [.126,.266]	.018 [$-.000$,.045]	.172 [.122,.235]
Train only	.046 [.008,.145]	.336 [.194,.495]	.046 [$-.045$,.156]	.526 [.368,.697]
SwissMetro only	.033 [.022,.053]	.013 [.010,.017]	.024 [.021,.027]	.013 [.010,.017]
Car only	.222 [.087,.371]	.367 [.213,.533]	.067 [.058,.076]	.034 [.005,.127]
Exclude Train	.163 [.063,.272]	.272 [.153,.391]	.056 [.048,.064]	.023 [.005,.092]
Exclude SwissMetro	.147 [.056,.251]	.282 [.188,.400]	.035 [.015,.054]	.174 [.110,.273]
Exclude Car	.017 [.007,.073]	.195 [.105,.286]	.021 [$-.023$,.083]	.307 [.203,.407]

Table 10: Frozen preference-learning protocol. Counts refer to baseline rows; each has three generated intervention rows.

Component	Frozen value
Split	Respondent-disjoint; seed 20260805; train/validation/test respondents 834/178/179; baseline rows 7,506/1,602/1,611
Human reference	Primary MNL estimated on training respondents only
Base policy	Qwen2.5-7B-Instruct; frozen local weights; float16 on Apple MPS
Choice protocol	Prompt terminates in `Answer:`; exact next-token probabilities for leading-space `A`, `B`, and `C`; unavailable alternatives masked before normalization
LoRA	Rank 16; $\alpha = 32$; dropout .05; target modules q, k, v, o, gate, up, and down projections
Optimization	One epoch; learning rate 10^{-5}; cosine schedule; .03 warmup; batch size 1; gradient accumulation 8; maximum length 256
SFT	Completion-only loss; 7,506 examples; 939 optimizer steps
DPO	Observed choice preferred to each available unchosen choice; 13,851 pairs; 1,732 steps; primary $\beta = .1$
Robustness matrix	Seeds 11, 22, and 33; near-step-matched SFT (1,736 versus 1,732 DPO steps); DPO $\beta \in \{.05, .1, .3\}$
Primary endpoint	Full fixed test cohort; $G^{\text{core}} = (G_{\text{time}} + G_{\text{fare}})/2$
Uncertainty	10,000 respondent-clustered bootstrap draws; separate 200-draw reference-refit analysis for human-model response gaps
Software	Python 3.13.15; PyTorch 2.8; Transformers 4.55.4; TRL .21; PEFT .17.1

Table 11: Core evidence after holding baseline behavior equal.

Study	Matched set	Metric	Result
Dose–response	824	Curve gap	Time .048 [.043,.053]; fare .031 [.027,.038]
Semantic paraphrase	679	Response gap	Time .050; fare .042; luggage .056
Cross-model	102	Signed response	Moonshot under-reacts; Kimi over-reacts
Cross-provider	200	Gap vs. repeat	Qwen .118/.227; DeepSeek .030/.139

Table 12: Independent-provider replication protocol. Both models receive the same frozen scenarios and intervention text.

Field	Qwen replication	DeepSeek replication
Model	Qwen3-8B	DeepSeek-v4-pro
Cases	450 scenarios; 1,800 calls each	
Conditions	baseline, exact repeat, time -20%, fare -20%	
Decoding	greedy; thinking off	temperature 0; thinking off
Output check	strict `Choice:` parser	strict JSON and choice validation
Successful calls	1,800/1,800	1,800/1,800
Common matched	200 scenarios from 156 respondents	
Uncertainty	10,000 respondent-clustered bootstrap draws	

C Detailed Experimental Protocol

C.1 Common falsification protocol

The protocol tests whether choice learning transfers to intervention response. Study 1 diagnoses response disagreement within baseline choice matches. Study 2 evaluates transfer from SFT and DPO on a fixed test population, then tests direct response supervision as a repair, controlling for intervention exposure. Repeat, paraphrase, reference, and provider checks assess alternative explanations. OPeRA separately compares expressed considerations within action matches; this supporting study does not test the primary intervention RQ.

C.2 Study 1: SwissMetro controlled response

SwissMetro contains 10,719 valid stated-preference observations from 1,191 respondents choosing among Train, SwissMetro, and Car. Each choice records alternative-specific time, fare, and headway plus traveler characteristics. Moonshot-v1-8k receives the original scenario and produces one choice at temperature zero. We then modify SwissMetro time, fare, or luggage while holding the remainder of the scenario fixed. The main matched cohort contains cases where the model reproduces the recorded human baseline choice.

Human counterfactual response is estimated with respondent-grouped, out-of-fold discrete-choice models. Our primary reference is a multinomial logit model with alternative-specific time, fare, and headway coefficients, alternative-specific constants, and prespecified interactions between traveler characteristics and mode utilities. A nonlinear reference adds piecewise-linear time and fare terms and interactions capturing observable preference heterogeneity. We compare both specifications with a pooled MNL baseline on identical respondent-disjoint folds. Before using either fitted model as a response reference, we require (i) lower held-out log loss than the pooled baseline and (ii) the expected signs for time and fare in every fold. Reporting results under both admissible references quantifies sensitivity to functional-form assumptions.

As a less model-dependent diagnostic, we exploit the nine scenarios per respondent. For time and fare separately, we retain each respondent's pair with the smallest standardized Euclidean distance over all nonfocal varying attributes, then keep the closest half of those pairs. We compare observed and fitted changes in SwissMetro choice using switch-direction agreement and Spearman correlation. Luggage is constant within respondent, so it has no analogous paired validation and remains a model-dependent heterogeneity estimate. These are population diagnostics, not individual causal estimands. Respondent-clustered resampling propagates repeated-choice dependence.

The multi-level extension evaluates SwissMetro time and fare at $\{-30, -20, -10, 0, 10, 20, 30\}\%$ changes from baseline. Its fixed sample contains all 224 matched SwissMetro choices plus 300 matched Train and 300 matched Car choices, poststratified to recover the baseline-choice composition of the 4,110-case matched cohort. The curve endpoint averages G_k over the six nonbaseline levels. Baseline and -20% conditions are repeated contemporaneously; agreement with their earlier versions diagnoses batch sensitivity. We additionally define the earlier-to-contemporaneous baseline change as an exact-repeat placebo. On the identical fixed sample, we compare this placebo with contemporaneous-baseline-to--20% time and fare changes. We report switching, signed SwissMetro-share response, and normalized L_1 response magnitude, with paired respondent-clustered intervals. This test asks whether an attribute intervention produces more change than ordinary re-elicitation; it does not assume that every intervention must exceed the placebo on every summary metric.

We add three safeguards. An exact-repeat condition measures response without a substantive change. A semantically equivalent prompt is evaluated on 824 matched cases. Kimi-k3 repeats baseline and interventions on a weighted 450-scenario sample, including a cohort matched jointly by the human and both LLMs. For the independent-provider replication, local Qwen3-8B and DeepSeek-v4-pro each receive the same frozen 450 scenarios under baseline, exact-repeat, time -20%, and fare -20% conditions, with deterministic decoding and thinking disabled. We report full, model-specific matched, and Qwen–DeepSeek common-matched cohorts, each constructed before resampling. These controls distinguish attribute response from generic output instability, prompt wording, and one-model or one-provider idiosyncrasy.

C.3 STUDY 2: PREFERENCE-LEARNING SEPARATION

Qwen2.5-7B-Instruct is evaluated under Base, SFT, and DPO. Respondents are disjoint across train (834 respondents; 7,506 observations), validation (178; 1,602), and test (179; 1,611). Neither objective sees counterfactual examples. SFT uses completion-only cross-entropy. DPO ranks the observed choice above each available unchosen alternative, yielding 13,851 pairs. Both use one epoch, LoRA rank 16, and learning rate 10^{-5}. The primary DPO setting uses $\beta = .1$ across three seeds; seed-22 sensitivity uses $\beta \in \{.05, .1, .3\}$.

The response-aware comparison uses the same respondent split and LoRA rank, one epoch, three seeds, and a learning rate of 3×10^{-6}. It augments baseline cross-entropy with finite-difference supervision on paired time and fare contexts generated from training respondents only; luggage is never used in this loss. To isolate the loss from exposure to perturbed inputs, the choice-augmentation ablation uses the identical 7,506 baseline/time/fare training triples, seed 22, optimizer schedule, 939 steps, and three forward passes, but replaces response-vector matching with ordinary hard-choice cross-entropy at each context. Thus its intended contrast changes the supervision target, not the intervention inputs or effective exposure.

The frozen behavioral endpoints are accuracy and NLL. We also report balanced accuracy, choice-share TVD, Brier score, confidence, and confidence when wrong. The response-alignment endpoint fixed before confirmatory scoring is

$$G_m^{\text{core}} = (G_{m,\text{time}} + G_{m,\text{fare}})/2 \tag{15}$$

on the full fixed test cohort. Time and fare enter because both have within-respondent validation and stable directions under both admissible human references; luggage is secondary because it is constant within respondent and its reference direction is specification-sensitive. Negative model-minus-Base changes indicate closer response alignment. Model contrasts use 10,000 respondent-clustered bootstrap samples; the reference-refitted sensitivity uses 200 draws.

C.4 STATISTICAL ESTIMANDS AND UNCERTAINTY

We report accuracy, NLL, Brier score, choice-share TVD, confidence, and G_k. The common estimand is a population response vector on a fixed cohort: trained models contribute exact answer-token probabilities, prompted APIs contribute elicited one-hot choices, and the human side contributes fitted choice probabilities. We never equate these population responses with individual potential outcomes. Model-comparison intervals use 10,000 respondent-clustered draws that preserve paired observations. Matched sets are fixed before resampling; cross-model comparisons additionally use a common matched cohort, whereas the training endpoint uses every fixed test case. Appendix A gives formulas, aggregation details, and secondary estimands.

Study 3 uses 225 exact human–model action matches from 604 eligible OPeRA shopping events. Randomized blind coding compares expressed consideration sets; protocol and label definitions are deferred to Appendix A because this exploratory study does not support the primary interventional claim.

D Detailed Results and Sensitivity Analyses

D.1 Human response benchmark

Both admissible references improve held-out log loss over pooled MNL. A within-respondent diagnostic retains 596 time and 596 fare pairs; across 256 observed switches, direction agreement is 64.5% (primary) and 66.0% (nonlinear), with pooled Spearman correlations .20 and .22. These moderate checks support population response ordering, not individual counterfactual truth. The evidentiary pattern is specification-stable where the paper makes its primary claim: both references pass the same ex ante sign and log-loss criteria, both support the time–fare discrepancy, and reference-refitted cluster intervals exclude zero for both attributes. Luggage is excluded from the core endpoint because its direction is not stable across specifications.

D.2 Behavior improves without response alignment

Across three seeds, SFT and DPO both raise accuracy from .443 to about .701 (Table 3). SFT reduces NLL to .669; DPO raises it to 3.270 and assigns at least .99 probability in 80.9–83.9% of cases.

Near-step-matched SFT (1,736 versus 1,732 DPO steps) yields .709 accuracy but a .042 core gap. At the raw-probability scale, every choice-only adapted condition improves time alignment, worsens fare alignment, and fails to improve the .030 Base core gap.

Validation-fitted temperature scaling reduces DPO NLL from 3.19–3.39 to .665–.683 and ECE from .265–.278 to .049–.058, but calibrated core gaps remain .036–.038 versus .029 for calibrated Base. Paired increases are .0092 [.0034,.0150], .0084 [.0022,.0145], and .0074 [.0016,.0131]. Calibration repairs confidence, not response structure at the primary β.

Across $\beta = .05, .1, .3$ at seed 22, raw NLL falls from 4.999 to 1.381 and the core gap from .070 to .052, but each raw core contrast remains above Base. After scaling, the $\beta = .3$ core increase is unresolved (.0046 [$-.0010$,.0103]), while its fare gap still increases by .0438 [.0338,.0536] and its time gap decreases. Thus β changes severity, not the attribute-selective rotation (Appendix 22).

Exploratorily, DPO raises SwissMetro probability by .144–.195 after a displayed fare reduction even for GA holders whose effective fare remains zero; no-GA gaps also persist (Appendix Table 23).

Response-aware supervision. $\lambda = 5$ attains a .0064 validation core gap but misses the frozen .703 accuracy floor. Fixed $\lambda = 1$ passes it and, across three test seeds, attains $.690 \pm .005$ accuracy and a $.0212 \pm .0011$ core gap, versus $.701 \pm .005$ and $.0358 \pm .0025$ for seed-matched SFT. The paired core change is $-.0146$ [$-.0168$,$-.0125$], while accuracy is unresolved and NLL worsens by .0302 [.0087,.0521]. Fare drives the gain; time is unresolved and held-out luggage does not improve (Appendix Table 6).

This improvement is not explained by exposure to intervention contexts alone. A seed-22 ablation applies ordinary hard-choice cross-entropy to the same baseline, time, and fare inputs, but removes the response-vector loss. Its test core gap is .0308, compared with .0224 for response-aware training; the paired difference is .0084 [.0076,.0093], while their accuracy difference is unresolved ($-.0062$ [$-.0199$,.0074]). The response-aware advantage is driven by fare (.0161 [.0147,.0175]); thus extra intervention data help relative to ordinary SFT, but do not recover the benefit of supervising how probabilities should move (Appendix Table 7).

D.3 Prompted-model response diagnostics

Moonshot completes all 10,719 baseline scenarios and 32,157 core interventions. Its unconditional behavior is not a strong human simulation: it over-selects Train and under-selects SwissMetro. Within baseline matches, the nonlinear human reference changes SwissMetro probability by $+6.52$, $+6.16$, and -5.14 points for time, fare, and luggage; corresponding model-minus-human gaps are -6.49 [-6.96,-6.03], -4.90 [-5.57,-4.24], and $+3.81$ [$+2.70$,$+4.90$].

Semantic reformulation preserves the conclusion. Of 824 original matched cases, 679 (82.4%) rematch; time/fare/luggage gaps remain .050/.042/.056. Exact repetition switches 1.0% [.5,1.7], while fare -20% switches 5.1% and raises SwissMetro share 1.3 points. Relative to placebo these

changes are 4.0 [2.5,5.7] and 1.4 [.4,2.5] points. Time increases switching by 1.6 [.5,2.8] points but not aggregate share reliably (Appendix Table 8).

Baseline matching is selective: Moonshot matches 38.3% overall but only 3.6% of SwissMetro choices; maximum standardized difference is 1.06 and weighting leaves limited overlap (Appendix Table 16). The matched analysis is therefore conditional, not population-representative.

Among 224 joint SwissMetro matches, fitted human probability rises 3.6/1.9 points under time/fare changes, while Moonshot share falls 13.0/9.8 points; GA holders correctly have near-zero human fare effects. Luggage remains secondary because its reference direction is specification-sensitive. Refitting the human reference in 200 respondent-clustered bootstrap draws preserves the core result: time-gap intervals are [0.031, 0.071] under the primary reference and [0.042, 0.076] under the nonlinear reference; fare-gap intervals are [0.019, 0.041] and [0.021, 0.047]. These intervals incorporate reference-estimation uncertainty conditional on each model class.

The multi-level extension confirms a response-geometry difference from -30% to $+30\%$, rather than an artifact of the original dose. Moonshot's time/fare slopes are $-.049/-.090$, versus $-.267/-.193$ for the primary reference; curve gaps are .048 [.043,.053] and .031 [.027,.038]. The nonlinear reference agrees, while earlier and contemporaneous anchors agree on at least 98.2% of weighted choices (full curves in Appendix F.1.1).

Cross-model tests show heterogeneous errors rather than universal under-reaction: Moonshot under-reacts to time and fare on common matches, whereas Kimi over-reacts and can reverse luggage direction. Independently served Qwen and DeepSeek models both retain fare mismatch on 200 common matches: gap-minus-repeat is .227 [.149,.321] and .139 [.088,.217]. Equal-choice weighting and leave-one-choice-out checks preserve this result (Appendix Table 9).

D.4 Supporting evidence from expressed considerations

Among 225/604 matched OPeRA actions, exact consideration-set agreement is 12.9% and Jaccard is .424, only .034 above an action/length-compatible null ($p = .0007$). User-disjoint incremental value is .0081 nats; with one coder, this limited same-action rationale signal remains exploratory.

D.5 Choice-learning control summary

Table 13: Choice-learning controls. Brackets are paired 95% respondent-clustered intervals; G denotes the time–fare core gap unless specified.

Test	Key result	Interpretation
Near-step-matched SFT	Accuracy .709; $G = .042$ vs. Base .030	Extra training does not close the gap.
Calibrated DPO, $\beta = .1$	NLL .665–.683; $G = .036$–$.038$ vs. calibrated Base .029	Confidence repair leaves response mismatch.
β sensitivity, seed 22	At calibrated $\beta = .3$, $\Delta G = .0046$ [$-.0010$,.0103]; fare $\Delta G_k = .0438$ [.0338,.0536]	Core contrast is unresolved; fare deterioration persists.

E SPORTS PRELIMINARY: EXPRESSED CONSIDERATIONS

Role in this paper. This preliminary motivated the distinction between matching an action and matching the considerations expressed with that action. It is not an interventional response study and supplies neither an estimate of internal cognition nor confirmatory evidence for the SwissMetro claims.

Data and protocol. Twenty basketball decision scenarios were completed voluntarily and without compensation by 24 informed participants, producing 480 human responses. The same scenarios elicited 500 model responses. Each response contained an action and a natural-language rationale. Rationales were mapped to eight multi-label dimensions: SPACE, TIME, TEAM, GOAL, ROLE, RISK, SKILL, and TACTIC. These labels describe expressed text content; they are not measurements of private mental states or faithful model reasoning traces.

Table 14: Descriptive action distributions in the sports preliminary.

Source	Action A	Action B	Action C	Action D
Human (%)	61.9	18.1	10.8	9.2
Model (%)	90.2	8.4	0.6	0.8

The action distributions already differ ($\chi^2 = 124.03$, $df = 3$, $p < .001$; Cramer's $V = .205$), so this sample is not presented as a high-fidelity behavioral simulation. Its useful observation is narrower: instances with the same action can still express different consideration sets, motivating the conditional and interventional tests in the main paper.

Table 15: Expressed-rationale label frequencies. Gap is model minus human; intervals are bootstrap 95% intervals from the exploratory analysis.

Label	Human	Model	Gap	95% interval
Time	.585	.840	+.255	[.195,.310]
Role	.581	.818	+.237	[.182,.294]
Skill	.831	.982	+.151	[.115,.186]
Space	.948	1.000	+.052	[.033,.073]
Goal	.992	1.000	+.008	[−.002,.023]
Risk	.821	.552	−.269	[−.325,−.214]
Team	.465	.168	−.297	[−.350,−.244]
Tactic	.419	.106	−.313	[−.368,−.263]

The model more frequently verbalizes execution-oriented labels and less frequently verbalizes team-, tactic-, and risk-related labels. These are stable differences in coded language, not evidence that the model intrinsically uses different cognitive mechanisms.

Prompt calibration. An exploratory prompt inserted the diagnosed label gaps and asked the model to express more human-like considerations. The mean absolute label-frequency gap increased from .187 to .255: two dimensions improved, five worsened, and one was unchanged. Thus changing requested rationale style can rotate expressed label frequencies without reliably moving them toward the human distribution. Because generated rationales need not faithfully report model computation, this result is retained only as motivation for the response-policy tests.

Limitations. The sample is small, the action distributions are not matched globally, and the analysis relies on an ontology applied to self-reports and generated text. The study has no controlled attribute interventions, respondent-disjoint prediction split, or independently replicated annotation. For these reasons, the main paper treats sports as a toy example and grounds its principal claims in SwissMetro response experiments instead.

Table 16: Baseline-match selection and limited-overlap sensitivity.

Quantity	Full cohort	Matched	IPW matched
SwissMetro choice share	0.580	0.055	0.432
Car choice share	0.287	0.651	0.386
Time switch rate	0.050	0.031	0.101
Fare switch rate	0.082	0.043	0.079
Luggage switch rate	0.054	0.034	0.218
Maximum absolute SMD	–	1.064	0.470

Baseline match rates are .850 for Train, .036 for SwissMetro, and .869 for Car. IPW uses inverse estimated match propensity, truncated at the 99th percentile. The effective weighted sample size is 711 and residual imbalance remains substantial; IPW is therefore an overlap stress test, not a corrected population estimate.

F Cross-Domain Decision-Skill Validation

The sports studies motivate the possibility of observational equivalence: the same action may be accompanied by different expressed reasons. We next replace rationale-only comparison with two tests grounded in original human behavior. SwissMetro supplies revealed choices and manipulable attributes; OPeRA supplies observed shopping actions and contemporaneous human rationales. The estimand is functional alignment of an elicited policy, not identity of unobserved biological or neural mechanisms.

F.1 SwissMetro: matched choice and counterfactual response

SwissMetro contains 10,719 stated-choice observations among Train, SwissMetro, and Car. For each original scenario, the LLM receives the observed traveler characteristics and available alternatives. Baseline fidelity is evaluated against the recorded choice. The primary response analysis then conditions on cases in which the human and LLM choose the same baseline alternative and applies controlled time, fare, and luggage interventions.

The human comparison is not a generated rationale. It is an out-of-fold discrete-choice benchmark estimated from the observed human data. For intervention k, let $\Delta\mathbf{p}_k^H$ and $\Delta\mathbf{p}_k^M$ denote human and model changes in the choice-share vector. The normalized structural gap is

$$G_k = \frac{1}{2}\left\|\Delta\mathbf{p}_k^M - \Delta\mathbf{p}_k^H\right\|_1. \tag{16}$$

Respondent-clustered bootstrap intervals preserve repeated observations. Exact-repeat prompts estimate ordinary elicitation instability, and a semantically equivalent template tests prompt dependence.

Five respondent-disjoint folds validate the human reference. The pooled, primary, and nonlinear specifications attain accuracies of 65.8%, 65.0%, and 65.5%, with log losses of 0.810, 0.769, and 0.756, respectively. Both richer references improve on the pooled loss and retain negative time and effective-fare slopes in every fold. A complementary nearest-pair diagnostic retains 596 within-respondent pairs for time and 596 for fare. Among 256 pairs with an observed switch into or out of SwissMetro, direction agreement is 64.5% for the primary and 66.0% for the nonlinear reference; pooled Spearman correlations are .20 and .22. Because luggage does not vary within respondent, its response remains a specification-sensitivity analysis rather than a paired revealed-choice check.

F.1.1 Multi-level dose–response extension

The single-dose analysis is extended to attribute changes $\mathcal{D} = \{-30, -20, -10, 0, 10, 20, 30\}\%$ for SwissMetro travel time and listed fare. This extension was designed after inspecting the original -20% result but before scoring the newly generated multi-level outputs; it is therefore a robustness extension rather than a preregistered endpoint. Baseline and -20% time/fare conditions are repeated contemporaneously on the same 824 cases and provide the primary anchors; the corresponding earlier outputs are retained only to quantify batch or model-version sensitivity. All levels use the identical traveler description, decoding temperature, and output format.

Table 17: Response-structure decomposition in the Moonshot baseline-matched cohort.

Intervention	No SM-status change	Reversal	Same-dir. over	G_k
Time	0.986	0.007	0.007	0.061
Fare	0.977	0.008	0.015	0.032
Luggage	0.978	0.007	0.015	0.033

"No SM-status change" means that the intervention does not move the deterministic LLM choice into or out of SwissMetro; Train–Car transitions are therefore not counted as an SM response. Categories compare each deterministic LLM change with the fitted nonlinear human-reference probability change. Because an LLM switch is discrete, the magnitude categories are diagnostic rather than causal estimands. Luggage is secondary because its fitted human-response direction changes across admissible specifications.

Table 18: Exploratory heterogeneity of matched-cohort response gaps under the nonlinear reference.

Group	Time gap	Fare gap	Luggage gap
Human baseline: Train	0.031 [0.026,0.041]	0.061 [0.042,0.081]	0.046 [0.036,0.056]
Human baseline: SwissMetro	0.165 [0.115,0.221]	0.118 [0.079,0.159]	0.352 [0.276,0.432]
Human baseline: Car	0.067 [0.065,0.069]	0.054 [0.051,0.057]	0.007 [0.003,0.011]
No GA pass	0.068 [0.064,0.072]	0.040 [0.034,0.046]	0.024 [0.018,0.032]
GA pass	0.028 [0.011,0.043]	0.006 [0.000,0.021]	0.087 [0.072,0.102]
Business/commute	0.046 [0.034,0.057]	0.015 [0.007,0.032]	0.063 [0.045,0.083]
Other purpose	0.066 [0.061,0.070]	0.038 [0.032,0.044]	0.024 [0.018,0.032]

Pointwise 95% respondent-clustered intervals; exploratory and not multiplicity-adjusted. Groups are evaluated within the fixed baseline-matched cohort.

The fixed evaluation sample contains all 224 baseline-matched observations in which the recorded and simulated choice is SwissMetro, plus seeded samples of 300 matched Train and 300 matched Car observations. Poststratification weights of 1.000, 4.030, and 8.923 for SwissMetro, Train, and Car recover the choice composition of the 4,110-case matched cohort. Results are reported both with these weights and by baseline-choice stratum; the oversampled SwissMetro stratum is not interpreted as population representative.

For attribute k, dose d, and policy m, let $\Delta\mathbf{p}_{m,k}(d) = \mathbf{p}_{m,k}(d) - \mathbf{p}_{m,k}(0)$. The curve-level response gap is

$$G_{m,k}^{\text{curve}} = \frac{1}{|\mathcal{D} \setminus \{0\}|} \sum_{d \in \mathcal{D} \setminus \{0\}} \frac{1}{2} \left\| \Delta\mathbf{p}_{m,k}(d) - \Delta\mathbf{p}_{H,k}(d) \right\|_1 . \tag{17}$$

Secondary diagnostics are the SwissMetro response-slope difference, aggregate direction agreement, transition flows, and the fraction of complete individual curves that violate monotonicity as time or fare worsens. The primary and nonlinear human references are fit on all valid observations and evaluated on the fixed weighted sample. Absolute fitted choice probabilities are not compared with deterministic LLM shares; all alignment measures are baseline-relative responses.

The transition decomposition shows why the mean gap is substantively informative. Within the 4,110 baseline-matched observations, Moonshot changes SwissMetro membership in only 1.4% of time interventions and 2.3% of fare interventions. Exploratory subgroup analysis finds the largest discrepancy in the 224 observations whose recorded and simulated baseline choice is SwissMetro: the nonlinear reference increases SwissMetro probability by 0.036 under the time intervention and 0.019 under the fare intervention, while Moonshot changes SwissMetro shares by -0.130 and -0.098. The corresponding normalized gaps are 0.165 (pointwise 95% CI: 0.115–0.221) and 0.118 (0.079–0.159). By contrast, among GA-pass holders the reference fare response is zero by construction of effective fare, and the model–human gap is 0.006 (0.000–0.021). This negative-control-like pattern indicates that the estimated reference responds to the economic meaning of an attribute, not merely to its numerical perturbation. These subgroup intervals are exploratory and not multiplicity-adjusted; the SwissMetro-baseline stratum is also strongly selected by matching. The luggage direction changes across the primary and nonlinear admissible references, so luggage cannot support the core claim and is reported only as functional-form sensitivity.

Moonshot-v1-8k completes the full 10,719-scenario baseline and 32,157 core interventions. Its baseline behavior is strongly biased toward Train. More importantly, even within baseline-matched

Table 19: Reference-refit uncertainty for matched-cohort response gaps.

Reference	Intervention	Mean G_k	95% refit interval
Primary	Time	0.055	[0.031, 0.071]
	Fare	0.029	[0.019, 0.041]
	Luggage	0.036	[0.024, 0.049]
Nonlinear	Time	0.062	[0.042, 0.076]
	Fare	0.033	[0.021, 0.047]
	Luggage	0.032	[0.020, 0.046]

Percentile intervals use 200 respondent-clustered draws and refit the human reference in every draw. They quantify sampling and reference-estimation uncertainty conditional on each model class; they are not individual causal-counterfactual intervals. Luggage gaps remain nonzero because G_k is unsigned, even though the signed human response changes across specifications.

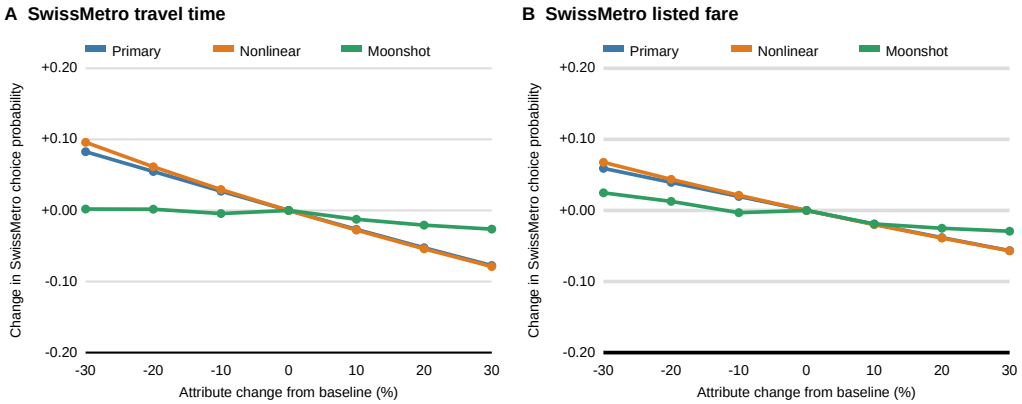

Figure 1: Baseline-relative SwissMetro dose–response curves on the fixed, poststratified matched sample. Attribute changes range from a 30% improvement to a 30% deterioration. Moonshot is substantially flatter than both human references, especially for travel time.

cases, its response to improved time and fare and changed luggage is smaller than the revealed-human response. In the semantic-template replication, 679 of 824 cases (82.4%) rematch the original baseline choice. Their normalized response gaps remain 0.050 for time, 0.042 for fare, and 0.056 for luggage. Thus the result is not explained by one wording of the task.

A Kimi-k3 replication improves population-level fit but does not eliminate the structural discrepancy. On a common cohort where the human, Moonshot, and Kimi agree at baseline, Moonshot under-reacts to time and fare while Kimi over-reacts; Kimi's luggage response can also move in the opposite direction from the human benchmark. Agreement at one observed point therefore does not identify the local policy around that point.

F.2 OPeRA: SAME SHOPPING ACTION, DIFFERENT CONSIDERATIONS

OPeRA links page context, interaction history, fine-grained shopping actions, and human rationales. Kimi-k3 produces a next action and a short rationale for 604 eligible events. All predictions are joined by unique instance identifier. The model matches the recorded action in 225 cases: 142 Scroll and 83 non-Scroll.

Human and model rationales for these matched actions are separated, randomized, and blind-coded over Price, Feature, Quality, Need, Risk, Social, Search, and Convenience. Multiple dimensions may be active. Across 225 pairs, mean normalized Hamming distance is 0.208, mean Jaccard similarity is 0.424, and exact consideration-set agreement is only 12.9%. A length- and action-compatible permutation null has mean Jaccard 0.390, giving a small but detectable alignment lift of 0.034 ($p = 0.0007$). The non-Scroll exact-match rate is 16.9%, so the divergence is not solely a consequence of the majority Scroll class.

Table 20: Multi-level response alignment. Lower curve gap is better.

Attribute	Human reference	Human slope	Moonshot slope	Curve gap [95% CI]
Time	Primary	-0.267	-0.049	.048 [.043, .053]
	Nonlinear	-0.290	-0.049	.052 [.047, .057]
Fare	Primary	-0.193	-0.090	.031 [.027, .038]
	Nonlinear	-0.207	-0.090	.033 [.028, .040]

Slopes regress the weighted SwissMetro share on the proportional attribute change across seven levels. Curve gap averages the normalized L_1 response gap across the six nonbaseline levels. Intervals use 10,000 respondent-clustered draws with the fitted human reference held fixed; reference-refit uncertainty is reported separately. Time-minus-fare gap intervals are [.010, .022] under the primary and [.012, .024] under the nonlinear reference. Old-versus-contemporaneous weighted agreement is 99.0% at baseline, 98.7% for time -20%, and 98.2% for fare -20%. Among 824 complete curves, 7.4% for time and 8.0% for fare violate individual monotonicity at least once. Because levels are elicited in separate API calls, these violations can include ordinary output instability and are diagnostic rather than the primary estimand.

We further ask whether the model-coded consideration vector predicts the human-coded vector after conditioning on the shared action and both rationale lengths. Five user-disjoint folds compare a logistic model containing action and log lengths with one that additionally contains the eight model labels. The mean held-out log-loss reduction is 0.0081 nats, compared with -0.0058 under 500 action- and model-length-bin-compatible permutations ($p = 0.002$, one sided). Gains are positive for Feature (0.0138), Quality (0.0030), Need (0.0366), and Social (0.0417), but negative for Price, Risk, Search, and Convenience. Hence model rationales carry a small amount of case-specific information beyond action and verbosity, but that information is sparse and dimension-specific.

These rationales are expressed considerations, not privileged reports of internal computation. Their value is conditional: after holding the action equal, they show that outcome agreement ordinarily coexists with different stated consideration sets. This complements the stronger SwissMetro counterfactual test, which evaluates policy behavior without requiring verbal self-explanation.

F.3 Synthesis

The sports ontology, SwissMetro interventions, and OPeRA matched-action study occupy different evidential levels. Sports rationales reveal structured differences in expressed concepts; OPeRA tests whether those differences remain after matching real actions; SwissMetro tests whether matched choices implement the same response function. Convergence across these tests supports the bounded conclusion that behavioral agreement is insufficient evidence of cognitive alignment. It does not establish access to latent neural mechanisms.

Table 21: Frozen test results for the completed robustness matrix ($N = 1{,}611$; 179 respondents). Lower is better except accuracy.

Policy	Accuracy	NLL	Brier	Time	Fare	Core
Base	.443	1.330	.759	.025	.034	.030
SFT seed 11	.701	.668	.403	.016	.051	.033
SFT seed 22	.705	.672	.402	.015	.061	.038
SFT seed 33	.696	.665	.400	.017	.054	.035
SFT step-matched, seed 22	.709	.673	.403	.012	.073	.042
DPO $\beta = .1$, seed 11	.708	3.229	.549	.004	.090	.047
DPO $\beta = .1$, seed 22	.693	3.194	.571	.005	.111	.058
DPO $\beta = .1$, seed 33	.703	3.389	.557	.003	.086	.044
DPO $\beta = .05$, seed 22	.694	4.999	.588	.012	.128	.070
DPO $\beta = .3$, seed 22	.677	1.381	.539	.009	.095	.052

F.4 Preference-learning protocol and robustness matrix

We adapt Qwen2.5-7B-Instruct using supervised fine-tuning (SFT) and Direct Preference Optimization (DPO). Respondents, rather than rows, are assigned to disjoint train, validation, and test sets using the frozen split seed 20260805. The split contains 834 training respondents (7,506 baseline observations), 178 validation respondents (1,602 observations), and 179 test respondents (1,611 observations). The corresponding intervention tables contain 22,518, 4,806, and 4,833 rows. The human response reference is fitted using training respondents only.

Neither training objective receives intervention examples. SFT uses completion-only cross-entropy over the observed choice. DPO ranks that choice above each available unchosen alternative, producing 13,851 pairs. The answer space consists of leading-space single-letter identifiers, and evaluation reads exact next-token probabilities after availability masking. This avoids confounding confidence with free-form answer length.

All adapters use one epoch, LoRA rank 16, LoRA $\alpha = 32$, dropout .05, and a 10^{-5} learning rate with cosine scheduling and .03 warmup. The primary DPO condition uses $\beta = .1$. SFT has 939 optimizer steps and DPO 1,732. A near-step-matched SFT control runs for 1,736 steps (four steps, or 0.23%, more than DPO). The robustness matrix additionally includes three seeds and $\beta \in \{.05, .1, .3\}$. All planned matrix conditions completed.

The primary response endpoint is

$$G_m^{\text{core}} = \frac{G_{m,\text{time}} + G_{m,\text{fare}}}{2} \tag{18}$$

on the full fixed test cohort. Negative model-minus-Base changes indicate closer response alignment. Intervals use 10,000 respondent-clustered bootstrap resamples.

Every complete adapted condition improves accuracy and time response alignment, but worsens fare alignment and the time–fare core endpoint. The near-step-matched SFT control shows that the pattern survives approximately equal optimizer exposure. All three completed DPO seeds are overconfident, with mean confidence .970–.976 and confidence when wrong .948–.956. These results support replicated attribute-selective transfer, not a general claim about DPO. Larger β improves probabilistic quality and attenuates the aggregate gap, but does not remove the opposing time and fare changes. Entries in Table 21 are point estimates.

F.5 Temperature-scaling control

For each condition, we fit one scalar temperature by minimizing NLL on the validation baseline, then freeze it for the test baseline and every test intervention. Thus test labels and intervention outcomes do not select the temperature. Scaling reduces DPO NLL to approximately the SFT level. At the primary $\beta = .1$, fare and core gaps remain larger than calibrated Base in all three seeds. At $\beta = .3$, the calibrated core contrast is unresolved, but its fare increase remains positive. Table 22 reports respondent-clustered intervals; accuracy is unchanged because scalar temperature preserves the choice ranking.

Table 22: Validation-fitted temperature scaling on the frozen test. Brackets are 95% respondent-clustered intervals.

Condition	Accuracy	NLL	Brier	Core gap	Fare gap
base_ts	0.443 [0.403, 0.481]	1.024 [1.001, 1.047]	0.635 [0.622, 0.649]	0.029 [0.026, 0.032]	0.016 [0.012, 0.020]
dpo_b005_seed22_ts	0.694 [0.656, 0.731]	0.693 [0.633, 0.756]	0.416 [0.376, 0.456]	0.040 [0.035, 0.044]	0.066 [0.056, 0.076]
dpo_b010_seed11_ts	0.708 [0.671, 0.744]	0.665 [0.602, 0.733]	0.398 [0.358, 0.439]	0.038 [0.034, 0.042]	0.065 [0.056, 0.073]
dpo_b010_seed22_ts	0.693 [0.655, 0.731]	0.683 [0.621, 0.748]	0.413 [0.373, 0.454]	0.037 [0.033, 0.041]	0.065 [0.057, 0.074]
dpo_b010_seed33_ts	0.703 [0.665, 0.739]	0.675 [0.612, 0.741]	0.405 [0.364, 0.447]	0.036 [0.032, 0.040]	0.057 [0.049, 0.065]
dpo_b030_seed22_ts	0.677 [0.639, 0.714]	0.691 [0.631, 0.755]	0.422 [0.382, 0.463]	0.033 [0.030, 0.037]	0.060 [0.052, 0.067]
sft_seed11_ts	0.701 [0.665, 0.737]	0.670 [0.604, 0.741]	0.404 [0.363, 0.446]	0.034 [0.030, 0.038]	0.052 [0.046, 0.059]
sft_seed22_ts	0.705 [0.667, 0.742]	0.665 [0.599, 0.735]	0.399 [0.359, 0.440]	0.037 [0.034, 0.041]	0.058 [0.051, 0.065]
sft_seed33_ts	0.696 [0.659, 0.732]	0.665 [0.601, 0.732]	0.400 [0.361, 0.440]	0.035 [0.032, 0.039]	0.054 [0.048, 0.060]
sft_stepmatched_seed22_ts	0.709 [0.672, 0.746]	0.661 [0.597, 0.729]	0.398 [0.358, 0.440]	0.041 [0.037, 0.045]	0.068 [0.060, 0.076]

F.6 Exploratory fare-gap heterogeneity

Table 23 separates GA-pass holders, whose effective public-transport fare is zero by construction, from other travelers. It reports pointwise respondent-clustered intervals without multiplicity adjustment. The DPO result is especially diagnostic: the model reacts to a lower displayed fare even when the prompt states that marginal fare remains zero. Positive gaps among nonholders show that the aggregate result is not driven only by this subgroup.

Table 23: Exploratory fare response gaps by GA-pass status.

Condition	No GA	GA pass	GA minus no GA
Base	.038 [.032,.044]	.011 [.008,.015]	$-.027$ [$-.034$,$-.020$]
SFT seed 11	.049 [.042,.057]	.061 [.047,.077]	+.012 [$-.005$,.029]
SFT seed 22	.060 [.052,.068]	.071 [.056,.087]	+.011 [$-.006$,.029]
SFT seed 33	.053 [.046,.060]	.060 [.049,.072]	+.007 [$-.006$,.021]
SFT step-matched 22	.066 [.057,.075]	.113 [.088,.139]	+.047 [.019,.075]
DPO seed 11	.079 [.060,.101]	.144 [.091,.202]	+.064 [.007,.125]
DPO seed 22	.095 [.074,.117]	.195 [.132,.263]	+.100 [.034,.171]
DPO seed 33	.074 [.055,.095]	.152 [.093,.218]	+.078 [.015,.147]

G PROOFS

This section collects the proofs of Propositions 1–3.

G.1 PROOF OF PROPOSITION 1 (SINGLE-CONTEXT NON-IDENTIFICATION)

Proof. Set the two baseline distributions equal. At $I_k(x)$, assign the human and model policies point masses on different available alternatives. Their baseline responses cancel in Eq. (6), leaving half the L_1 distance between the two intervened distributions, which equals one. □

G.2 PROOF OF PROPOSITION 2 (CHOICE RISK DOES NOT DETERMINE RESPONSE RISK)

Proof sketch. Consider a binary choice with one observed context x and an intervened context $I(x)$ outside observational support. Let the human distribution at x be $(1/2, 1/2)$, and let it assign probability one to the first alternative at $I(x)$. Construct p_n with baseline distribution $(1/2+\epsilon_n, 1/2-\epsilon_n)$, where $\epsilon_n > 0$ tends to zero, and with probability one on the second alternative at $I(x)$. Under log loss, the excess observational risk is positive and tends to zero. The response gap, however, tends to one. Hence no bound on response risk that vanishes with positive excess choice risk can hold for all such policies. □

G.3 PROOF OF PROPOSITION 3 (ENDPOINT-ERROR BOUND)

Proof. Expand the response vectors in Eq. (6) and apply the triangle inequality to the baseline and intervened distribution errors. □

H EXPERIMENTAL DESIGN AND RESULTS: SUPPORTING TABLES

This section collects the workflow, experimental specifications, and detailed comparisons supporting Sections **??** and 6.

Training settings. SFT and DPO use one epoch, rank-16 LoRA, learning rate 10^{-5}, and seeds 11, 22, and 33. DPO uses 13,851 pairs with primary $\beta = 0.1$. Response-aware training retains the split, rank, epoch count, and seeds, with learning rate 3×10^{-6}. Its loss weight is selected using validation core gap subject to the accuracy floor. The seed-22 augmentation control matches intervention inputs, 939 steps, and three forward passes.

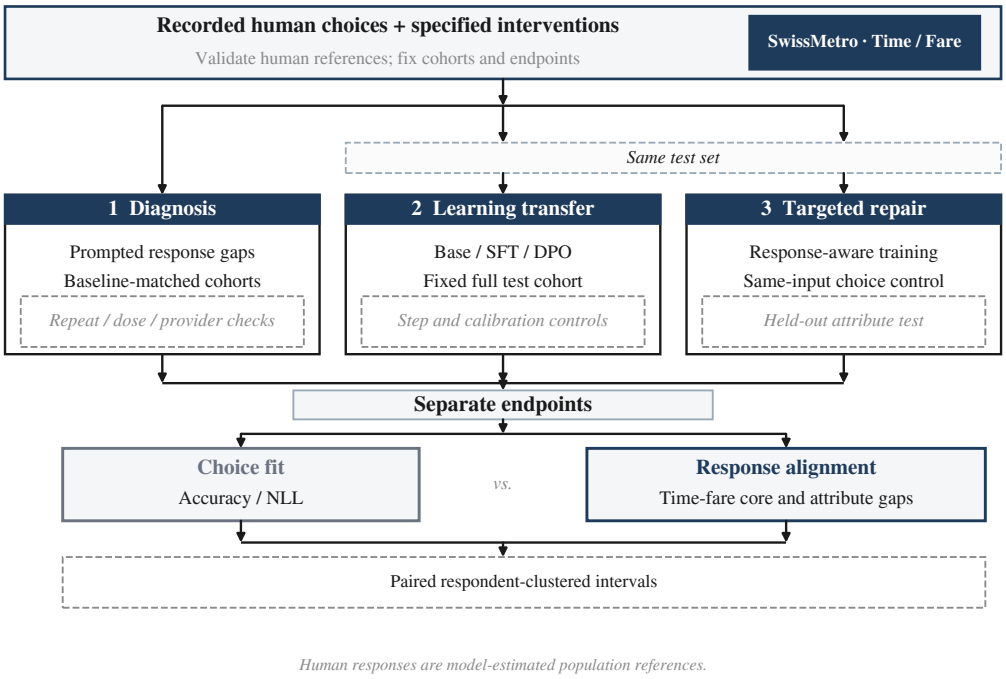

Figure 2: From matching choices to evaluating responses. We first examine intervention responses within baseline-choice matches, then compare choice-only and response-aware training on a fixed full test cohort. Robustness controls accompany these comparisons to assess alternative explanations for the observed gaps. Scenes and displayed choices are schematic.

Table 24: The comparisons behind the experimental design. SwissMetro follows response alignment from prompted choices through training; OPeRA offers a complementary view of expressed considerations.

Experiment	Sample and comparison
Prompted diagnosis	Moonshot baseline choices and time, fare, and luggage interventions; responses evaluated after baseline-choice matching. The dose sample has 824 cases and six nonzero levels from -30% to $+30\%$.
Choice-learning transfer	Qwen2.5-7B base, SFT, and DPO policies, with three training seeds. Train/validation/test: 7,506/1,602/1,611 choices from 834/178/179 respondents. All policies use the same full test cohort.
Response supervision	Response-aware training on baseline/time/fare triples from training respondents, compared with SFT and same-input hard-choice augmentation. Luggage is withheld from response supervision.
OPeRA support	604 events; blind-coded consideration sets compared within exact action matches. One coder, an action/length-compatible null, and a non-Scroll analysis support an exploratory assessment of expressed alignment.

Table 25: Interpreting a response gap. Each control asks whether the comparison is sensitive to a particular aspect of elicitation, evaluation, or training.

Question	Comparison
Does elicitation matter?	Exact-repeat elicitation and an equivalent prompt template on 824 cases.
Do dose and reference matter?	Seven intervention levels including baseline; two admissible human reference models.
Does it extend across models?	Kimi, plus Qwen3-8B and DeepSeek evaluated on 450 frozen cases each.
Does matching affect the gap?	Full, own-matched, and common-matched cohorts; balance checks and reweighting.
Is input exposure sufficient?	Near-step-matched SFT and hard-choice augmentation using the same intervention inputs.
Do confidence or settings matter?	Validation-only temperature fitting; $\beta \in \{0.05, 0.1, 0.3\}$ at seed 22.
Does alignment transfer?	Luggage interventions, excluded from response supervision.

Table 26: Human-reference validation on respondent-disjoint folds.

Reference	Accuracy ↑	Log loss ↓	Time/fare signs
Pooled MNL	.658	.810	5/5
Primary	.650	.769	5/5
Nonlinear	.655	.756	5/5

Arrows indicate the preferred direction; signs are checked in all five folds.