XChoice: Explainable Evaluation of AI–Human Alignment in LLM-based Constrained Choice Decision Making

Weihong Qi, Fan Huang , Rasika Muralidharan , Jisun An , Haewoon Kwak
Indiana University Bloomington
Correspondence: {wq3, hwkwak}@iu.edu

Abstract

We present XCHOICE, an explainable framework for evaluating AI-human alignment in constrained decision making. Moving beyond outcome agreement such as accuracy and F1 score, XCHOICE fits a mechanism-based decision model to human data and LLM-generated decisions, recovering interpretable parameters that capture the relative importance of decision factors, constraint sensitivity, and implied trade-offs. Alignment is assessed by comparing these parameter vectors across models, options, and subgroups. We demonstrate XCHOICE on Americans' daily time allocation using the American Time Use Survey (ATUS) as human ground truth, revealing heterogeneous alignment across models and activities and salient misalignment concentrated in subgroups defined by race (Black respondents) and marital status (Spouse: Present). We further validate robustness of XCHOICE via an invariance analysis and evaluate targeted mitigation with a retrieval-augmented generation (RAG) intervention. Overall, XCHOICE provides mechanism-based metrics that diagnose misalignment beyond surface outcome matching and make the effects of targeted interventions measurable at the mechanism level. Code and data access instructions are available at https://github.com/WeihongQ/xchoice.

1 Introduction

As large language models (LLMs) are increasingly used to assist, predict, or simulate human decisions, understanding when and how their behavior aligns with human decision making has become a central concern. Recent work has explored LLM-driven systems for multi-entity interaction and strategic reasoning, as well as the dynamics of LLM–LLM and LLM–human interactions (Xi et al., 2025; Ruan et al., 2023; Wang et al., 2024; Piatti et al., 2024; Duan et al., 2024; Zhang et al., 2024a). Despite this progress, we still lack systematic metrics that explain *why* models align or misalign with humans: which decision-relevant attributes and constraints they consider, how those factors are weighted, and how trade-offs are resolved.

Constrained decision making offers a natural setting to study these questions. Individuals allocate limited resources such as time, budget, or capacity across competing options and make attribute-based trade-offs under feasibility constraints (de Bekker-Grob et al., 2012; Brock and Durlauf, 2001; Train, 2009; Berry, 1994). Examples include grocery shopping under a budget, course selection under time constraints, or allocating hours across work, sleep, and leisure. Such decisions are ubiquitous and increasingly relevant to LLM applications in recommendation, forecasting, and behavior simulation (Lyu et al., 2024; Jia et al., 2024; Park et al., 2023; Dillion et al., 2023; Xie et al., 2024; Li et al., 2024), and mis-specified trade-offs can propagate to downstream outcomes and reduce the reliability of decision-support systems. Constrained decision making is also central to modern LLM agents themselves, where tool-using and planning pipelines operate under explicit feasibility limits such as bounded tool calls, context length, or budget caps, making a mechanism-level diagnostic broadly applicable to evaluating LLM behavior, not only to modeling human decisions.

Most existing alignment evaluations in choice settings focus on outcome agreement, using metrics such as accuracy or F1 score (Qi et al., 2025; Zhang et al., 2024b), aggregate replication (Hua et al., 2023), or option distributions (Horton, 2023). These measures do not reveal whether models and humans rely on similar decision mechanisms, nor identify which decision factors, and the subgroups defined by them, account for observed discrepancies.

We propose XCHOICE, an explainable framework for evaluating AI-human alignment through estimated decision mechanisms in constrained de-

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3833–3861
November 6-10, 2026 ©2026 Association for Computational Linguistics

Table 1: Comparison of three alignment evaluation families. XCHOICE recovers interpretable latent preference weights via an explicit choice mechanism, and supports generalization to decisions under new environments

Aspect	Outcome Metrics	Reduced-Form	XCHOICE
Measurement	Match rate on decisions (AI vs. human)	Associations between inputs and observed choices	Latent preference weights (trade-offs) in a utility model
Explainability	Low (Black-box)	Moderate (Correlational)	High (Mechanistic / Causal)
Constraints	Not applicable	Must be imposed explicitly	Built in via model structure
Generalizability	Not generalize beyond observed outcomes	Sensitive to distribution shifts	Transfers to new environments; stable preference weights

cision making. XCHOICE fits a mechanism-based decision model to human and LLM-generated decisions, recovering interpretable parameters that capture decision-factor weights, constraint sensitivity, and implied trade-offs; alignment is assessed by comparing these parameters and resulting decision patterns to enable diagnosis and the evaluation of targeted interventions. Table 1 summarizes how XCHOICE improves on outcome-based metrics and reduced-form analyses through mechanism recovery, fine-grained diagnostics, and robustness and intervention evaluations.

Our work makes three contributions. First, we formalize alignment in constrained decision making as a mechanism-comparison problem rather than outcome matching. Second, we introduce structural, interpretable metrics that identify where and how LLM decision mechanisms diverge from human baselines. Third, using daily time allocation decisions from the American Time Use Survey (ATUS) (Flood et al., 2025), we demonstrate how XCHOICE diagnoses salient misalignment and evaluates targeted interventions at the mechanism level.

2 Related Work

2.1 Structural Estimation in Constrained Decision Making

The Lucas critique argues that correlations between observed variables and choices are fragile indicators of behavior, since reduced-form relationships can shift when policies or constraints change (Lucas, 1976; Sims, 1980; Kaplan, 2009). Structural estimation addresses this by estimating deep parameters that capture stable preferences, trade-off weights, and constraints, supporting counterfactual analysis and mechanism-based interpretation (McFadden, 1974, 1981; Rust, 1987; Keane and Wolpin, 1997). This invariance motivates the widespread use of structural models in labor economics, industrial organization, and household decision making (Berry, 1994; Berry et al., 1995; Eckstein and Wolpin, 1989; Train, 2009). For AI–human alignment, this perspective shifts the focus from matching observable outcomes to whether an AI system captures the structure of human decision making, providing a more robust basis for assessing whether models generalize human-like behavior across environments.

2.2 LLM-based Decision Making and Evaluation

Recent work has explored LLMs' ability to make human-like decisions, particularly under discrete or constrained settings (Eigner and Händler, 2024; Sun et al., 2025). Benchmarks such as Temporal Constraint-based Planning, PlanningArena, and asynchronous planning environments evaluate time-management and scheduling problems via constraint satisfaction, task completion, and efficiency relative to optimal solutions (Ding et al., 2025; Zheng et al., 2025); budget-constrained and participatory budgeting tasks assess resource allocation via feasibility, regret, and utility loss (Zheng et al., 2024).

These evaluations focus primarily on aligning human and LLM outputs (Xiao and Wang, 2025; Jung et al., 2025), leaving decision mechanisms opaque: models may satisfy performance criteria while employing unstable or socially implausible reasoning. This gap is particularly salient in socially situated domains like time management and resource allocation, where the underlying rationale is as critical as the choice itself, motivating methods that recover decision motivations for interpretable insights into LLM behavior.

3 Method

We present the three-step XCHOICE framework for mechanism-based AI-human alignment evaluation: we first model decisions as constrained optimization, then estimate interpretable parameters from human and LLM decisions, and lastly compare the resulting weights and decision pat-

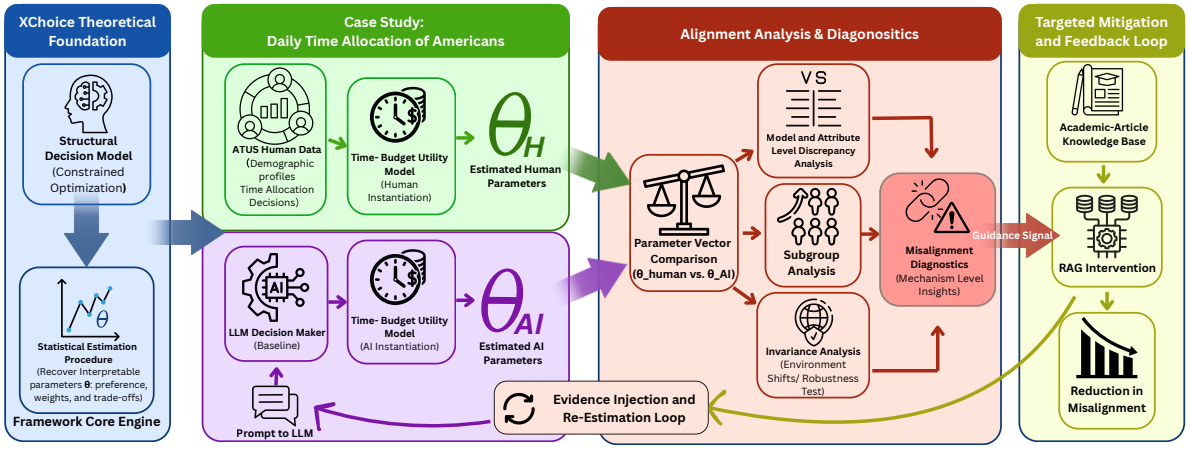


Figure 1: End-to-End Pipeline for XCHOICE Evaluation with ATUS Data

terns to diagnose misalignment and guide targeted interventions. Figure 1 summarizes the end-to-end workflow, illustrated with the ATUS instantiation.

3.1 Constrained Optimization Model

We model individual decision making as a constrained optimization problem. Each individual m chooses a vector x by solving:

$$\begin{aligned} &\max_{x \in D} \; f_m(x; \theta_d) \\ &\text{s.t.} \quad g_i(x) = c_i, \quad h_j(x) \geq d_j, \end{aligned} \tag{1}$$

where $f_m(\cdot)$ is the objective function, θ_d captures trade-off parameters, and g_i, h_j represent equality and inequality constraints, respectively.

Solving this optimization yields the individual's optimal decision rule

$$x_m^* = x_m^*(c_i, d_j; \theta_d), \tag{2}$$

where $x_m^* \in X^*$ is determined by trade-off parameters and constraints, and X^* is the set of optimal decisions across individuals. To analyze AI-human alignment at the group level, we aggregate individual decisions via an operator $\mathcal{F}(\cdot)$:

$$D = \mathcal{F}(X^* \mid c, d; \theta_d). \tag{3}$$

D is the aggregated decision by subgroup, enabling alignment to be examined by age, income, or race, and helping reveal systematic differences in how humans and LLM-based decision makers allocate resources, resolve trade-offs, or exhibit biases.

3.2 Statistical Estimation of Parameters

To recover the parameters θ_d of $f_m(x; \theta_d)$, we employ *inverse optimization* (Ahuja and Orlin, 2001; Keshavarz et al., 2011), which estimates parameters by identifying the values that render the observed choices x_m^* optimal given the constraints and context. Formally, inverse optimization minimizes $\|x_{\text{obs}} - x^*(\theta)\|^2$, where $x^*(\theta)$ is the model-implied optimal decision. When $x^*(\theta)$ is closed-form, as in our ATUS instantiation in Section 4, this reduces to a Non-Linear Least Squares problem; for discrete decisions, it maps to likelihood-based discrete-choice estimation. The full algorithmic specification and a discrete-choice instantiation are provided in Appendices B and J. We apply this procedure identically to human and LLM decisions, so that comparing the resulting estimates evaluates alignment based on underlying trade-off mechanisms rather than surface-level outcomes.

3.3 Evaluating Alignment: Metrics, Robustness, and Intervention

The estimated parameters $\hat{\theta}$ summarize the decision process by quantifying how decision-relevant attributes shift the relative attractiveness of each option under the constraint, so differences in $\hat{\theta}$ indicate where an LLM reweights factors and alters implied trade-offs compared with humans. We use these estimates to diagnose model- and attribute-specific misalignment, assess robustness via invariance tests, and guide targeted interventions.

Alignment Metrics. We quantify AI–human alignment by comparing each model's estimated parameter vector $\hat{\theta}_l$ with the human baseline $\hat{\theta}_H$ using *model-level* and *attribute-level* analyses. At the model level, we report cosine similarity $CosSim(\hat{\theta}_H, \hat{\theta}_l)$ to measure whether the model matches the human directional pattern of relative weights, and we summarize overall discrepancy magnitude by first defining feature-wise deviations $\Delta_{l,f} = |\theta_{l,f} - \theta_{H,f}|$ and then averaging them as $M_l = \frac{1}{F}\sum_{f=1}^{F} \Delta_{l,f}$ on all F features. At the at-

tribute level, we aggregate the same deviations across L models as $A_f = \frac{1}{L}\sum_{l=1}^{L}\Delta_{l,f}$ to identify features whose weights systematically depart from the human baseline, where larger A_f indicates stronger and more consistent misalignment. To avoid over-interpreting noisy estimates from small subgroups, we report the within-group count N_g alongside A_f and restrict headline conclusions to well-powered features with $N_g \geq 50$; bootstrap percentile intervals over respondents are provided in Appendix I.5.

Structural Robustness. To ensure the alignment assessment reflects stable mechanisms rather than superficial correlations, we conduct an **invariance analysis**. Motivated by the Lucas critique (Lucas, 1976), we shift the distribution of observed covariates and measure how much $\hat{\theta}$ changes, benchmarking against reduced-form baselines that are expected to drift under environmental shifts. We use three metrics: MAD (mean absolute coefficient change), RelL2 (relative change in ℓ_2 magnitude), and $1 - CosSim$ (directional rotation). Formal definitions and counter-factual data shift details are in Appendix G.

Targeted Mitigation. Finally, XCHOICE supports a diagnose-then-intervene loop. Using A_f to localize misalignment, we inject decision-relevant information to target the corresponding trade-off distortions, instantiated with a retrieval-augmented generation (RAG) intervention. We then re-estimate parameters from post-intervention decisions to quantify changes in alignment, whether improvements or regressions.

4 Case Study: Time Allocation of Americans

We apply XCHOICE to a concrete decision task: the allocation of time across daily activities. We formalize the decision structure, estimate parameters from human and LLM-generated decisions, and compare the inferred mechanisms to identify sources of alignment and discrepancy.

4.1 Dataset

We estimate model parameters using the American Time Use Survey (ATUS) (Flood et al., 2025), a nationally representative dataset with rich demographic and time-use measures. We partition the 24-hour budget into four categories: work, leisure, sleep and personal care, and other activities, with average daily allocations of 247, 266, 578, and 349 minutes respectively and substantial cross-individual dispersion indicating heterogeneous trade-offs. Individual features include standardized age, education, and weekly earnings, plus binary indicators for sex, spouse status, and race. Our final sample contains 4,307 individuals; preprocessing and summary statistics are in Appendices C and D.

4.2 Modeling Optimal Time Allocation

We model daily time allocation as a continuous-choice problem in which individuals distribute a fixed amount of time across four discrete activity categories $j \in \{L, W, S, O\}$: leisure, work, sleep and personal care, and other activities. Each individual i chooses an allocation vector h_i that maximizes utility:

$$\max_{\{h_{ij}\}} U_i(h_i;\theta) = \sum_{j\in\{L,W,S,O\}} \exp(\theta_j^\top X_i)\cdot \ln(h_{ij}) \tag{4}$$

subject to the time-budget constraint $\sum_j h_{ij} = T$, where $T = 1440$ mins. Although representing a day as $T = 24$ hours is more natural, we set $T = 1440$ minutes to match the ATUS time unit and to allow finer-grained variation in the generated allocations. The vector X_i contains demographic and socioeconomic characteristics affecting how the individual values each activity.

Solving this problem yields $h^*_{ij} = T\frac{\exp(\theta_j^\top X_i)}{\sum_k \exp(\theta_k^\top X_i)}$. For empirical estimation, we work with time shares $s_{ij} = h^*_{ij}/T = \frac{\exp(\theta_j^\top X_i)}{\sum_k \exp(\theta_k^\top X_i)}$, where the parameters θ_j represent the relative importance of attributes in determining time allocation.

4.3 Parameter Estimation

Statistical Method. We estimate $\theta = (\theta_L, \theta_W, \theta_S, \theta_O)$ by fitting model-predicted shares to observed time-use data. Let s^{obs}_{ij} denote the observed time share for activity j and individual i. Because the model is nonlinear in θ, we use Non-Linear Least Squares:

$$\min_\theta \sum_i \sum_j \left(s^{\text{obs}}_{ij} - \frac{\exp(\theta_j^\top X_i)}{\sum_k \exp(\theta_k^\top X_i)} \right)^2 .$$

The estimated parameters provide an interpretable representation of how individuals trade off activities, conditioned on their demographic attributes.

Reference activity and interpretation. Because shares depend on the *relative* values of $\theta_j^\top X_i$ across activities, parameters are identified up to a normalization. We use *Other Activities* as the reference by setting $\theta_O = \mathbf{0}$ and report coefficients for *Leisure*, *Work*, and *Sleep and Personal Care* relative to this baseline. Coefficients then capture how features shift the propensity for each focal activity versus *Other* with the daily time budget fixed, and differences in $\hat{\theta}_j$ ($j \in L, W, S$) reflect how humans and LLMs reweight trade-offs relative to the same reference. We verify in Appendix I.1 that the diagnostic ranking is not an artifact of this choice: re-estimating with $\theta_S = \mathbf{0}$ yields an identical Top-5 set with Spearman rank correlation 0.879.

Instantiating decision makers. We generate synthetic decision makers by embedding individual ATUS attributes such as age, education, and income into a standardized prompt template (Appendix A). This conditions the LLMs to produce time-allocation decisions reflecting realistic demographic profiles.

LLMs under evaluation. We evaluate a diverse set of LLMs spanning closed- and open-source families: GPT-4o, Claude-3.7, Qwen-2.5, Llama-3.3, and DeepSeek-V3. Model versions, endpoints, and implementation details are in Appendix E.

5 Experiments: Misalignment Analysis and RAG-Based Mitigation

5.1 Misalignment Analysis

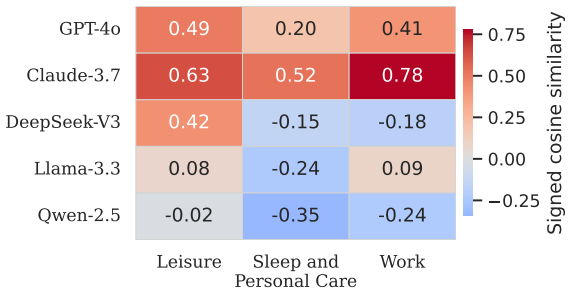


Figure 2: Cosine Similarity Between LLM and Human Estimated Parameter Vectors by Activity

Model-Level Misalignment Analysis. Figure 2 reports model-level alignment via cosine similarity $CosSim(\hat{\theta}_l, \hat{\theta}_H)$ for each activity-specific parameter vector. The heatmap shows strong activity dependence: Claude-3.7 is consistently best aligned, especially for *Work* (0.78), while GPT-4o is moderately positive across activities. Several open-source models show mixed or sign-reversing alignment. For example, DeepSeek-V3 is positive for *Leisure* (0.42) but negative for *Sleep and Personal Care* and *Work*, and Qwen-2.5 performs worst with negative similarity in all three activities. These negative values indicate systematic differences in attribute weighting, consistent with mismatched priors about time-use patterns or difficulty capturing the constrained trade-offs in the human data of LLMs.

Table 2: Model-level alignment deviation between humans and LLM-based models. For each model l, M_l is the mean absolute deviation of its coefficients from the human baseline (lower is better); "Share" reports the fraction of total deviation $\sum_l M_l$ accounted for by each model.

Model	Mean M_l	Share (%)
Qwen-2.5	0.189	26.4
Llama-3.3	0.180	25.1
GPT-4o	0.128	17.8
DeepSeek-V3	0.126	17.6
Claude-3.7	0.094	13.1

Complementing cosine similarity, Table 2 reports the model-level divergence score M_l, which captures the *magnitude* of coefficient gaps from the human baseline. The ranking is broadly consistent with the cosine-based results: Claude-3.7 has the smallest deviation ($M_l = 0.094$), while DeepSeek-V3 (0.126) and GPT-4o (0.128) show moderate gaps. In contrast, Llama-3.3 (0.180) and Qwen-2.5 (0.189) deviate the most and together account for over half of the total divergence, indicating systematically larger departures from human trade-off weights.

Table 3: Feature-wise alignment deviation between human and LLM-based models. For each feature f, A_f is the mean absolute deviation of model coefficients from the human baseline; "Share" reports the fraction of total deviation $\sum_f A_f$ accounted for by each feature.

Feature	Mean A_f	Share (%)
Race: Pacific Islander	0.374	26.0
Race: Native American	0.223	15.6
Race: Black	0.194	13.5
Spouse: Present	0.171	11.9
Sex	0.098	6.8
Age	0.095	6.6
Earnings	0.076	5.3
Race: Asian	0.072	5.0
Spouse: Unmarried	0.070	4.9
Education	0.062	4.3

Attribute-Level Misalignment Analysis. To further decompose and identify the sources of mis-

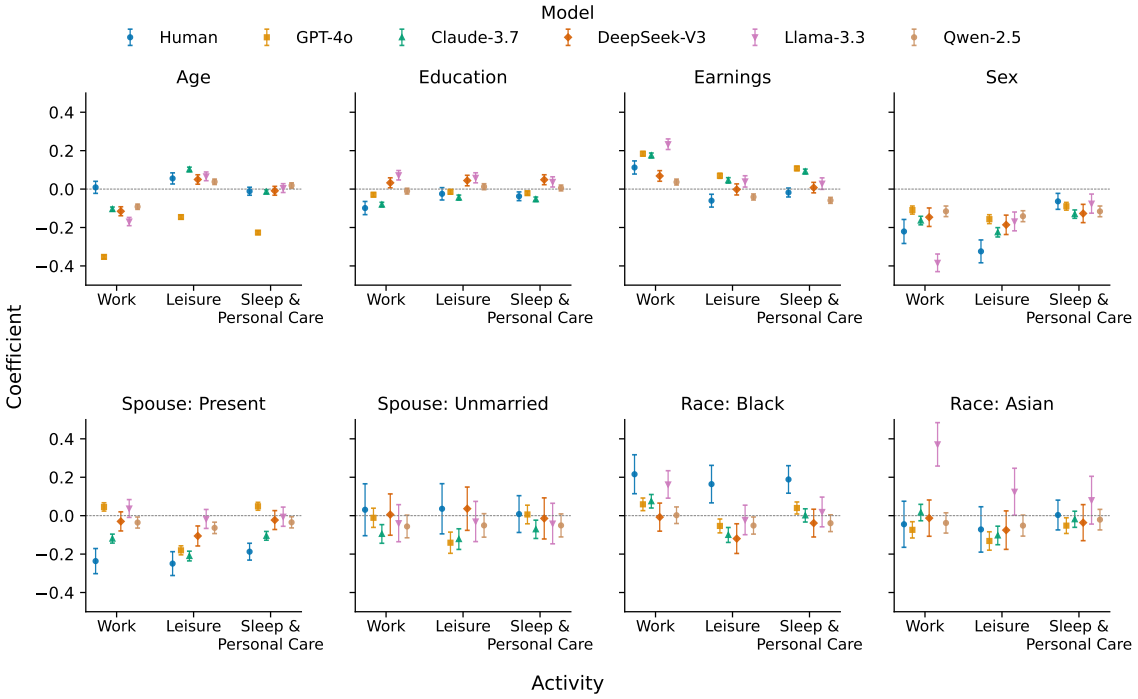


Figure 3: Attribute-Level Comparison of Estimated Decision Weights Between Humans and LLMs

Table 4: Misalignment Analysis: Groups with Largest Deviation from Human Baseline

Activity	Group	Human	GPT-4o	Claude-3.7	Deepseek-V3	Llama-3.3	Qwen-2.5
Work	Race: Black	**0.216**±0.101	**0.059**±0.033	**0.075**±0.035	-0.008±0.073	**0.163**±0.071	0.002±0.043
Work	Spouse: Present	**-0.236**±0.065	**0.046**±0.022	**-0.119**±0.023	-0.030±0.050	0.037±0.046	**-0.036**±0.029
Leisure	Race: Black	**0.164**±0.098	**-0.053**±0.036	**-0.100**±0.039	**-0.119**±0.077	-0.022±0.077	**-0.051**±0.045
Leisure	Spouse: Present	**-0.249**±0.062	**-0.180**±0.023	**-0.210**±0.025	**-0.106**±0.052	-0.017±0.049	**-0.063**±0.030
Sleep/PC	Race: Black	**0.189**±0.072	**0.040**±0.031	0.001±0.034	-0.039±0.072	0.019±0.077	-0.039±0.044
Sleep/PC	Spouse: Present	**-0.187**±0.044	**0.049**±0.021	**-0.105**±0.023	-0.023±0.049	-0.005±0.050	**-0.035**±0.029

Notes: Sleep/PC denotes *Sleep and Personal Care*. Entries are estimated coefficients from the constrained time-allocation model, reported as estimate ± 95% confidence interval. Coefficients are interpreted *relative to the reference activity* ("Other") under the normalization $\theta_O = \mathbf{0}$: a positive coefficient means the subgroup tends to allocate *more* time to the listed activity (vs. "Other") than the reference group; a negative coefficient means *less* time. We compare coefficients across humans and models to diagnose which decision factors are over- or under-weighted by each LLM. **Bold** indicates statistical significance ($p < 0.05$). Underline marks, within each activity, the model with the largest absolute deviation from the human coefficient for that group (worst alignment).

alignment between human and LLM decision makers, Figure 3 compares the estimated coefficients feature by feature (with complete estimates in Table 8 in Appendix F). We focus on (i) whether models recover the same *direction* and statistical salience of each weight and (ii) the *magnitude* of divergence from the human baseline. For readability, we omit the estimation results on *Race: Pacific Islander* and *Race: Native American* as subfigures because the underlying sample sizes are too small to support stable inference, even though these features rank highly in the deviation summary as shown in Table 3[1]. Across most covariates, the results do not reveal a single consistent alignment/misalignment pattern: models may track the human sign and significance for some attributes in one activity while diverging in another, consistent with the activity-dependent heterogeneity observed in Figure 2.

Despite this overall heterogeneity, Figure 3 and Table 3 highlight two attributes that exhibit particularly large and consistent deviations from the human baseline across models: *Race: Black* and *Spouse: Present*. Table 4 reports the largest deviation cases within each activity and motivates our focus on these two groups in the main text, while the full coefficient estimates are provided in Table 8 Appendix F. For *Race: Black*, the human coefficients are positive and statistically significant across all three activities (Leisure: 0.164, Sleep and Personal Care: 0.189, Work: 0.216), indicating that relative to the reference category

[1]The results with *Race: Pacific Islander* and *Race: Native American* are shown in Figure 5 in Appendix F

Other, Black respondents allocate systematically more time to each focal activity. In contrast, most LLMs substantially attenuate these effects and often reverse the sign, with the largest discrepancies occurring in Leisure (DeepSeek-V3: -0.119, Claude-3.7: -0.100, GPT-4o: -0.053), suggesting that the models encode qualitatively different trade-offs for this group.

A similar pattern holds for *Spouse: Present*. In the human baseline, this coefficient is negative and significant across Leisure (-0.249), Sleep and Personal Care (-0.187), and Work (-0.236), implying that married individuals shift time away from these activities toward the reference category. Several models shrink these negative effects sharply toward zero, and some nearly eliminate them. For example, GPT-4o is close to zero for *Work* and *Sleep and Personal Care*, producing the largest deviations highlighted in Table 4.

Under the structural interpretation of coefficients as attribute-dependent reweighting of constrained trade-offs, these discrepancies indicate that LLMs encode systematically different priorities for race and marital status when allocating a fixed time budget. These mechanism-level gaps are not apparent from outcome agreement alone, but become salient when alignment is evaluated through estimated decision weights. Formal Wald tests on the 30 human–LLM coefficient differences for these two attributes support this reading: 27 of 30 remain significant after Holm correction (Appendix F).

5.2 Validating XChoice against Reduced-Form Regression

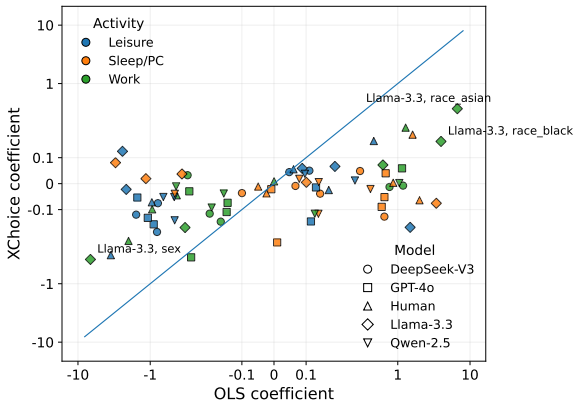


Figure 4: Coefficient Scatter Plot Comparing Reduced-Form Regression and XCHOICE Estimates

Comparison to Reduced-Form Regressions. Figure 4 compares reduced-form linear regression (OLS) coefficients (x-axis) with XCHOICE mechanism-based coefficients (y-axis) for the same model–activity–feature tuples, with detailed OLS estimation results reported in Table 9 in Appendix F. If the two approaches yielded similar estimates, points would concentrate around the $45°$ line; instead, we observe substantial dispersion and a clear "shrinkage" pattern: many coefficients that appear extreme under OLS map to much smaller XCHOICE estimates, with points pulled toward zero on the y-axis. This contrast is especially visible for several demographic features in the *Work* category, where OLS produces large-magnitude effects (far right/left), while XCHOICE yields more moderate values that reflect trade-offs under the daily time-budget constraint. The most prominent outliers, such as Llama-3.3 on *Race: Asian* and *Race: Black*, highlight cases where OLS suggests strong associations but the structural mechanism implies weaker or qualitatively different reweighting once substitution across activities is accounted for. Overall, the figure suggests that reduced-form regressions can amplify activity-specific correlations, whereas XCHOICE produces more conservative, mechanism-consistent weights that are less driven by marginal correlations and more aligned with constrained allocation structure.

Table 5: Parameter drift under mild covariate shifts (lower is better). Bold indicates the smaller drift within each shift (XCHOICE vs. OLS).

Shift	Method	MAD	RelL2	1-CosSim
Earnings lift	XCHOICE	**0.0045**	**0.0251**	**0.0003**
	OLS	0.0380	0.0432	0.0008
Age band lift	XCHOICE	**0.0012**	**0.0055**	**0.0000**
	OLS	0.0169	0.0240	0.0003
Race mix shift	XCHOICE	**0.0374**	**0.1487**	**0.0111**
	OLS	0.4716	0.7522	0.1881
Spouse mix shift	XCHOICE	**0.0205**	**0.1088**	**0.0059**
	OLS	0.2190	0.3724	0.0692

Robustness/Invariance Analysis. To further assess the robustness of XCHOICE under mild data and policy drift, we conduct an invariance analysis by introducing small shifts in the distributions of earnings, age, race, and spouse status (data shift details in Appendix G). Table 5 reports the invariance analysis results with three different metrics. Each row reports the drift between $\theta^{(0)}$ (pre-shift estimates) and $\theta^{(1)}$ (post-shift estimates) under one shift; smaller values indicate stronger invariance. Across all four shifts, the structural estimator exhibits consistently lower MAD/RelL2 and near-zero $1 - \mathrm{CosSim}$ relative to OLS, implying that

Table 6: RAG mitigation results. Cosine similarity to the human baseline.

Model	Subgroup	Baseline	RAG	Δ
GPT-4o	Race: Black	-0.827	-0.782	+0.045
	Spouse: Present	0.276	0.741	+0.465
DeepSeek-V3	Race: Black	-0.774	-0.673	+0.101
	Spouse: Present	0.704	-0.015	-0.719

the *direction* of the estimated preference vector (i.e., relative trade-offs across covariates) is largely preserved under mild environmental perturbations. This supports the interpretation that XCHOICE estimation better captures stable latent "value weights," whereas OLS coefficients are more entangled with sample-specific correlational structure and thus drift more under covariate reweighting.

Beyond the environmental shifts above, we run a suite of additional robustness checks that perturb the modeling and identification choices themselves: an equivalent reference-category re-normalization (Appendix I.1), removal of one plausibly non-core covariate (Appendix I.2), ε smoothing of corner solutions (Appendix I.3), a CES-style elasticity sweep (Appendix I.4), a non-parametric bootstrap for sampling uncertainty (Appendix I.5), a round-number granularity check (Appendix I.6), and a diary-day reweighting that matches the human estimand to the typical day elicited from LLMs (Appendix I.7). The Top-5 most misaligned features are stable across all checks, with rank correlations of 0.88 or above, indicating that the diagnostic conclusions are not driven by any single modeling assumption.

5.3 Mitigating AI-Human Misalignment with RAG

Guided by the diagnostics, we focus mitigation on the two most consistently divergent attributes, *Race: Black* and *Spouse: Present*, and target GPT-4o and DeepSeek-V3 as representative closed- and open-source models. As shown in Table 4, they are the only models with worst-case misalignment in two activities across the two focal subgroups. We then apply a RAG intervention that conditions LLM decisions on evidence retrieved—via embedding similarity—from an academic-article knowledge base built from domain-specific papers, containing 43 and 50 instances on time allocation by marital status and race, respectively, and re-estimate parameters to measure changes in mechanism-level alignment. Implementation details are provided in Appendix H.

Table 6 summarizes how the RAG intervention changes mechanism-level alignment for the two targeted subgroups. Baseline and post-intervention parameters are estimated on the same random subsample of $N = 431$ respondents, and uncertainty is quantified by paired bootstrap resampling of respondents. The intervention yields one statistically credible improvement: for GPT-4o on *Spouse: Present*, cosine similarity to the human baseline rises from 0.37 to 0.84 ($\Delta = +0.47$, 95% CI $[0.09, 0.95]$). The remaining cells are flat (DeepSeek-V3 on *Race: Black*) or degrade (GPT-4o on *Race: Black* and DeepSeek-V3 on *Spouse: Present*); the degradations are directionally consistent but imprecise at this subsample size, and the two *Race: Black* cells rest on only $N_g = 42$ respondents, below our $N_g \geq 50$ threshold, so we treat them as exploratory.

Evidence direction vs. quoted statistics. Many retrieved passages quote explicit statistics (68% of retrieved contexts overall and 93% in the marriage condition), so a post-RAG shift could in principle reflect imitation of the quoted numbers rather than an adjustment of trade-off weights. We therefore re-run the intervention with retrieval, prompts, and decoding held fixed, replacing every quantitative figure in the retrieved passages with a qualitative expression of the same directional finding (Appendix H.7). The GPT-4o *Spouse: Present* gain is almost fully retained without statistics (0.82 vs. 0.84, i.e., 96% of the gain; between-arm difference 95% CI $[-0.13, 0.11]$), indicating that the shift is driven by the direction of the evidence rather than by replication of its numbers.

When does discrepancy-guided intervention help or hurt? The results point to three failure modes. *(i) Persona–evidence mismatch.* The two GPT-4o cells start from nearly identical baselines (0.37) yet diverge sharply: the marriage knowledge base, whose evidence concerns adult households, produces the credible gain, whereas the race knowledge base produces a large degradation, and its retrieved evidence includes findings on adolescent populations retrieved for adult respondents (Appendix H.5). Evidence relevance, rather than the size of the baseline gap, is the operative factor. *(ii) Intervening on already-aligned mechanisms.* DeepSeek-V3 on *Spouse: Present* starts close to the human pattern (0.92) and falls to 0.09 after RAG, suggesting that interventions should be gated on the diagnostic signal A_f rather than applied uni-

formly. *(iii) Evidence format.* In the same cell, the qualitative arm restores most of the loss (0.82; between-arm difference 95% CI $[0.19, 1.34]$), so quoted statistics can distort trade-offs that a model already encodes in a human-consistent way, even though genuine gains do not require them. These failures are visible only because alignment is re-estimated at the mechanism level after the intervention; we therefore view the RAG study as a proof of concept of the diagnose-then-intervene loop rather than a finished mitigation method.

6 Conclusions and Discussions

We propose XCHOICE, an explainable framework for evaluating AI-human alignment in constrained decision making via mechanism-based structural metrics. Rather than relying on outcome agreement alone, XCHOICE recovers and compares interpretable decision parameters that capture how humans and LLMs weight attributes, satisfy constraints, and resolve trade-offs. In our ATUS case study, model alignment is highly heterogeneous across activities and features, with salient discrepancies concentrated in subgroups such as *Race: Black* and *Spouse: Present* that are difficult to diagnose using accuracy-style metrics.

Beyond diagnosis, XCHOICE supports informed mitigation: in a proof-of-concept RAG intervention on GPT-4o and DeepSeek-V3 guided by the discrepancy signals, retrieved evidence that is relevant to the persona improves alignment, and the gain persists when quoted statistics are removed, whereas evidence that is mismatched to the persona or that targets an already human-consistent mechanism can degrade alignment, failures that post-intervention re-estimation makes directly visible. An invariance analysis under environment shifts further shows that the recovered $\hat{\theta}$ is more stable than reduced-form baselines, consistent with the Lucas-critique motivation. Beside continuous-share allocation, a logistic-choice instantiation on Anthropic HH-RLHF (Appendix J) recovers an interpretable preference vector over thirteen text features with significant directional signals on length, formatting, politeness, and refusal markers, indicating that XCHOICE extends naturally to discrete-preference settings. Overall, alignment hinges on whether models learn human-like trade-off weights that encode stable priorities, not just whether they match observed choices in a single context.

Limitations

Scope: descriptive, not normative. Throughout this paper, "misalignment" refers to a *descriptive* quantity, namely the behavioral divergence between an LLM and an observational human baseline under a shared feasibility constraint. It does not imply that the human baseline is normatively desirable, nor that closer alignment is always preferable. We therefore view XCHOICE as a diagnostic instrument, a "thermometer" that surfaces where decision mechanisms diverge and makes implicit trade-offs auditable, while normative judgments about which trade-offs are desirable remain downstream and application-dependent. This distinction is particularly important for analyses involving sensitive attributes such as race and marital status: identifying that a model encodes systematically different trade-off weights for a subgroup is informative for understanding model behavior, but does not by itself dictate how the model should behave.

Methodological scope. XCHOICE is designed for *constrained decision making* problems where choices can be formalized with an explicit objective, a feasible set, and interpretable trade-offs. While this class of decisions is common and consequential, LLMs are deployed in a much broader range of settings such as open-ended generation, multi-step reasoning, negotiation, and interactive assistance, that may not admit a clean constrained-optimization representation. Extending mechanism-based alignment metrics to those contexts remains an open direction. Concretely, applying XCHOICE requires observed decisions, decision-relevant covariates, and a specifiable feasible set. Such data are common (e.g., discrete choice experiments in health economics, consumer purchase panels, recommendation logs, and LLM agent trajectories under explicit budgets), and the HH-RLHF instantiation in Appendix J shows that covariates can also be computed from the choice objects themselves.

Estimation assumptions. Our approach inherits its limitations from structural estimation. The recovered parameters depend on the assumed model form and normalization choices, and they summarize behavior within the scope of observed covariates and constraints. Misspecification, omitted variables, or heterogeneity beyond what the model captures can bias estimates and complicate interpretation. Moreover, invariance is best viewed as an

empirical property rather than a guarantee: structural parameters may still drift under large regime shifts or when preferences and institutions evolve. These limitations motivate future work on more flexible structural formulations, richer sources of behavioral variation, and broader validation across decision environments.

Human benchmark and elicitation. ATUS records an individual-level diary for a single day, so household joint decisions (e.g., spouse behavior) and constraints such as commuting or caregiving are only partially reflected in the covariates. Because human and LLM estimates condition on the same covariate vector, XCHOICE compares implied trade-offs given identical information, but unobserved constraints shape what the human benchmark represents. In addition, LLMs are prompted for a typical day, whereas ATUS oversamples weekend diaries. We align the two estimands by reweighting (Appendix I.7), but LLM outputs elicited without day information are day-invariant by construction and cannot reproduce the weekday/weekend structure of human time use.

Mitigation study. The RAG intervention is a proof of concept covering two models, two subgroups, and a subsample of respondents, and the *Race: Black* cells are underpowered. Its outcome depends strongly on whether retrieved evidence matches the persona, so we do not present it as a general-purpose mitigation method.

Risks of misuse. Although XCHOICE is intended as an evaluation tool, mechanism-based diagnostics could be misused to engineer models that better imitate human decision patterns for manipulative or discriminatory purposes, especially when applied to sensitive attributes or subgroups. Subgroup-level analyses may also be over-interpreted as normative claims about how groups *should* behave, or may reinforce stereotypes if reported without appropriate context and uncertainty. To mitigate these risks, we report uncertainty alongside subgroup estimates and recommend avoiding deployment decisions based solely on subgroup-specific coefficients, without complementary qualitative and domain review.

Acknowledgments

This research was supported by the Republic of Korea's MSIT (Ministry of Science and ICT), under the Global Research Support Program in the Digital Field Program (RS-2024-00425354) supervised by the IITP (Institute of Information and Communications Technology Planning & Evaluation).

References

Ravindra K. Ahuja and James B. Orlin. 2001. Inverse optimization. *Operations Research*, 49(5):771–783.

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

Steven T. Berry. 1994. Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, 25(2):242–262.

William A. Brock and Steven N. Durlauf. 2001. Discrete choice with social interactions. *The Review of Economic Studies*, 68(2):235–260.

Esther W. de Bekker-Grob, Mandy Ryan, and Karen Gerard. 2012. Discrete choice experiments in health economics: A review of the literature. *Health Economics*, 21(2):145–172.

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

Zifeng Ding, Sikuan Yan, Moy Yuan, Xianglong Hu, Fangru Lin, and Andreas Vlachos. 2025. TCP: A benchmark for temporal constraint-based planning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22452–22475, Suzhou, China. Association for Computational Linguistics.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. 2024. GTBench: Uncovering the strategic reasoning capabilities of LLMs via game-theoretic evaluations. In *Advances in Neural Information Processing Systems*, volume 37, pages 28219–28253.

Zvi Eckstein and Kenneth I. Wolpin. 1989. The specification and estimation of dynamic stochastic discrete choice models: A survey. *The Journal of Human Resources*, 24(4):562–598.

Eva Eigner and Thorsten Händler. 2024. Determinants of LLM-assisted decision-making. *Preprint*, arXiv:2402.17385.

Sarah M. Flood, Liana C. Sayer, and Daniel Backman. 2025. American Time Use Survey Data Extract Builder: Version 3.3 [dataset]. College Park, MD: University of Maryland and Minneapolis, MN: IPUMS. https://doi.org/10.18128/D060.V3.3.

John J. Horton. 2023. Large language models as simulated economic agents: What can we learn from Homo Silicus? Working Paper 31122, National Bureau of Economic Research.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. 2023. War and peace (WarAgent): Large language model-based multi-agent simulation of world wars. *Preprint*, arXiv:2311.17227.

Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. 2024. GPT4MTS: Prompt-based large language model for multimodal time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23343–23351.

Jaehun Jung, Faeze Brahman, and Yejin Choi. 2025. Trust or escalate: LLM judges with provable guarantees for human agreement. In *The Thirteenth International Conference on Learning Representations*.

David Kaplan. 2009. *Structural Equation Modeling: Foundations and Extensions*, second edition, volume 10 of *Advanced Quantitative Techniques in the Social Sciences*. SAGE Publications, Thousand Oaks, CA.

Michael P. Keane and Kenneth I. Wolpin. 1997. The career decisions of young men. *Journal of Political Economy*, 105(3):473–522.

Arezou Keshavarz, Yang Wang, and Stephen Boyd. 2011. Imputing a convex objective function. In *2011 IEEE International Symposium on Intelligent Control*, pages 613–619, Denver, CO, USA. IEEE.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024. EconAgent: Large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15523–15536, Bangkok, Thailand. Association for Computational Linguistics.

Robert E. Lucas, Jr. 1976. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1:19–46.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024. LLM-Rec: Personalized recommendation via prompting large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 583–612, Mexico City, Mexico. Association for Computational Linguistics.

Daniel McFadden. 1974. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York.

Daniel McFadden. 1981. Econometric models of probabilistic choice. In Charles F. Manski and Daniel McFadden, editors, *Structural Analysis of Discrete Data with Econometric Applications*, pages 198–272. MIT Press, Cambridge, MA.

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

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In *Advances in Neural Information Processing Systems*, volume 37, pages 111715–111759.

Weihong Qi, Hanjia Lyu, and Jiebo Luo. 2025. Representation bias in political sample simulations with large language models. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1264–1267. ACM.

Jingqing Ruan, Yihong Chen, Bin Zhang, Zhiwei Xu, Tianpeng Bao, Guoqing Du, Shiwei Shi, Hangyu Mao, Ziyue Li, Xingyu Zeng, and Rui Zhao. 2023. TPTU: Task planning and tool usage of large language model-based AI agents. In *Foundation Models for Decision Making Workshop at NeurIPS 2023*.

John Rust. 1987. Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica*, 55(5):999–1033.

Christopher A. Sims. 1980. Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Chuanneng Sun, Songjun Huang, and Dario Pompili. 2025. LLM-based multi-agent decision-making: Challenges and future directions. *IEEE Robotics and Automation Letters*, 10(6):5681–5688.

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

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4):20:1–20:38.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, and 9 others. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

Feng Xiao and XiaoTian Wang. 2025. Evaluating the ability of large language models to predict human social decisions. *Scientific Reports*, 15(1):32290.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip H. S. Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behavior? In *Advances in Neural Information Processing Systems*, volume 37, pages 15674–15729.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024a. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14544–14607, Bangkok, Thailand. Association for Computational Linguistics.

Xinnong Zhang, Jiayu Lin, Libo Sun, Weihong Qi, Yihang Yang, Yue Chen, Hanjia Lyu, Xinyi Mou, Siming Chen, Jiebo Luo, Xuanjing Huang, Shiping Tang, and Zhongyu Wei. 2024b. ElectionSim: Massive population election simulation powered by large language model driven agents. *Preprint*, arXiv:2410.20746.

Yuanhang Zheng, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. 2024. Budget-constrained tool learning with planning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9039–9052, Bangkok, Thailand. Association for Computational Linguistics.

Zihan Zheng, Tianle Cui, Chuwen Xie, Jiahui Pan, Qianglong Chen, and Lewei He. 2025. PlanningArena: A modular benchmark for multidimensional evaluation of planning and tool learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31047–31086, Vienna, Austria. Association for Computational Linguistics.

Appendix

A Prompts to LLMs

System Instruction: You are an American making daily time allocation decisions across three activities: work, leisure, and other. Your goal is to maximize overall happiness within the 1,440 minutes available each day.

User Prompt Template: You are a {Gender Identity}, {Age} years old, ethnically identified as {Race}. Your highest level of education is {Educational Level}, and your weekly income is {Income}. Based on this background, how would you allocate your time across Work, Leisure, Sleep and Personal Care, and Other in a typical day? Please answer in the format: [Work, Leisure, Sleep and Personal Care, Other], using numbers to indicate minutes.

Listing 1: Prompt template for personalized LLM-based agents.

B Algorithmic Details of XCHOICE

This appendix provides the full algorithmic specification of the XCHOICE estimation pipeline summarized in Section 3. The core idea is that inverse optimization defines a single framework-level objective,

$$\min_{\theta} \mathcal{L}(\theta) = \left\| x_{\text{obs}} - x^*(\theta) \right\|^2, \tag{5}$$

while the appropriate numerical solver depends on the structure of $x^*(\theta)$. When $x^*(\theta)$ admits a closed form, as in the Cobb–Douglas time-budget instantiation of Section 4, Eq. (5) reduces directly to a Non-Linear Least Squares problem in which the closed-form $x^*(\theta)$ enters the objective and θ is obtained by standard numerical optimization. When $x^*(\theta)$ has no closed form, $x^*(\theta)$ is computed numerically inside the optimization loop while the objective itself remains unchanged. For discrete decisions, the same objective maps naturally to standard discrete-choice estimators: $x^*(\theta)$ is replaced by the choice probability $P(y \mid x; \theta)$, and θ is recovered by maximum likelihood. We provide a concrete discrete-choice instantiation on Anthropic HH-RLHF in Appendix J. Algorithm 1 summarizes the three-step pipeline used throughout the paper.

C Data Preprocessing

This section describes the data preprocessing procedures used to transform the raw survey data into a structured format suitable for subsequent analysis. The original ATUS extract contains 8,548 respondents. We first restrict the sample to individuals with non-missing information on key covariates, dropping observations coded as "not in universe" or with missing earnings (e.g., EARNWEEK = 99999.99). After these restrictions, the final analytic sample consists of 4,307 respondents.

We then construct a set of harmonized covariates. Race is recoded from the original multicategory RACE variable into four single-race indicators (Black, Native American, Asian, and Pacific Islander), using White-only respondents as the reference group. Educational attainment is collapsed from the detailed EDUCYRS codes into an ordered four-level variable (edu) distinguishing high school or less, some college, bachelor's degree, and any

Table 7: Summary statistics for attributes and time-use variables (N = 4,307).

Variable	Mean	Std. Dev.	Min	25th pctl	Median	75th pctl	Max
Panel A: Continuous attributes (standardized)							
Age	0.000	1.000	-2.054	-0.827	-0.077	0.809	2.718
Education (edu)	0.000	1.000	-1.347	-1.347	0.498	0.498	1.420
Weekly earnings (EARNWEEK)	0.000	1.000	-1.621	-0.765	-0.195	0.637	1.945
Panel B: Binary and indicator attributes							
Male	0.507	0.500	0.000		–		1.000
Spouse present	0.542	0.498	0.000		–		1.000
Unmarried partner present	0.058	0.233	0.000		–		1.000
Race: Black	0.107	0.309	0.000		–		1.000
Race: Native American	0.008	0.089	0.000		–		1.000
Race: Asian	0.068	0.252	0.000		–		1.000
Race: Pacific Islander	0.002	0.043	0.000		–		1.000
Panel C: Time-use variables (minutes per day)							
Work	246.97	260.66	0.00	0.00	135.00	488.00	1390.00
Leisure	266.33	197.41	0.00	120.00	225.00	385.00	1237.00
Sleep and personal care	578.16	136.48	0.00	500.00	570.00	645.00	1415.00
Other	348.53	210.83	0.00	185.00	305.00	480.00	1413.00

Note: Continuous predictors in Panel A are standardized to have mean zero and unit variance. For binary and indicator variables in Panel B, we report mean, standard deviation, minimum, and maximum; quantiles are not reported (indicated by "–").

Algorithm 1 XCHOICE: Mechanism Recovery and Alignment Diagnosis

Require: Observed decisions $x_{\text{obs}}^{(H)}$ and $x_{\text{obs}}^{(l)}$ from human and LLM l; constraints $\{g_i, h_j\}$; covariates X

Ensure: Estimated parameters $\hat{\theta}_H, \hat{\theta}_l$ and alignment diagnostics

Step 1: Forward solve. For a candidate θ, compute the model-implied optimal decision

$$x^*(\theta) \;=\; \arg\max_{x \in \mathcal{D}} \; f(x;\theta) \quad \text{s.t. } g_i(x) = c_i,\; h_j(x) \geq d_j.$$

Step 2: Inverse recovery. For each decision maker $\mathcal{M} \in \{H, l\}$, solve

$$\hat{\theta}_{\mathcal{M}} \;=\; \arg\min_{\theta} \; \| \, x_{\text{obs}}^{(\mathcal{M})} - x^*(\theta) \, \|^2.$$

Use NLLS with the closed-form $x^*(\theta)$ when it is analytic, as in the continuous-share case. Otherwise compute $x^*(\theta)$ numerically inside the optimization loop, or use a likelihood-based estimator $\max_\theta \sum \log P(y \mid x; \theta)$ for discrete choices.

Step 3: Diagnostic comparison. Compare $\hat{\theta}_H$ and $\hat{\theta}_l$ via the model-level metrics CosSim and M_l, and the attribute-level metric A_f, as defined in Section 3.3.

postgraduate education. Sex is recoded into a binary indicator for male (with female as the reference category), and household partnership status is represented by two indicators for spouse present and unmarried partner present (with no spouse or partner as the reference category).

In addition to these attributes, we retain respondents' time allocation across four activity categories—work, leisure, sleep and personal care, and other—each measured in minutes per day. The resulting dataset provides a clean and interpretable set of variables for the subsequent empirical analysis and LLM-based simulations.

Sex. Respondents' sex is recorded in the variable SEX (01 = male, 02 = female, 99 = NIU). We exclude observations coded as NIU (SEX = 99) and recode this variable into a binary indicator. Specifically, we define a variable *Male* that takes the value 1 if the respondent is coded as male (SEX = 01) and 0 if coded as female (SEX = 02), treating female as the reference category. This binary indicator is used as the sex covariate in the subsequent analysis.

Race. The original dataset reports respondents' race using the variable RACE, a multi-category code that distinguishes single-race and multiracial combinations (e.g., 0100 = White only, 0110 = Black only, 0120 = American Indian/Alaskan Native, 0131 = Asian only, 0132 = Hawaiian/Pacific Islander only, and codes $\geq$ 0200 for various multiracial combinations). To focus on clearly defined single-race groups and avoid small, heterogeneous multiracial cells, we restricted the sample to respondents with a single reported race by dropping all observations with codes greater than or equal to 200 (`drop if race >= 200` in Stata). We then used "White only" as the baseline category and constructed four binary indicator variables—*Race: Black*, *Race: Native American*, *Race: Asian*, and *Race: Pacific Islander*—which take the value 1 if the respondent belongs to the corresponding single-race group and 0 otherwise. These indica-

tors enter the subsequent models as race covariates, with White as the omitted reference group.

Income. Weekly earnings are measured using the variable EARNWEEK, which reports respondents' usual weekly income. In the original data, missing values are coded as 99999.99. We treat this code as missing and drop all such observations from the analysis (EARNWEEK = 99999.99). The resulting non-missing values of EARNWEEK are used as the weekly income covariate in our models.

Education. The survey records respondents' educational attainment using the variable EDUCYRS, which distinguishes detailed schooling levels (e.g., grades 1–12, one to four years of college, bachelor's degree, and various graduate degrees). For ease of interpretation and to avoid sparsely populated categories, we collapsed EDUCYRS into an ordered four-level variable edu. Specifically, we coded edu = 1 for respondents with at most a high school education (EDUCYRS $\leq$ 112, grades 1–12), edu = 2 for those with some college but no bachelor's degree ($112 <$ EDUCYRS ≤ 216), edu = 3 for respondents with a bachelor's degree (EDUCYRS $= 217$), and edu = 4 for respondents with any postgraduate education (EDUCYRS > 300), including master's, professional, and doctoral degrees. This four-level variable is used as the education covariate in subsequent analyses.

Spouse/partner status. We measure household partnership status using the variable SPOUSEPRES, which indicates whether a spouse or unmarried partner is present in the household (01 = spouse present, 02 = unmarried partner present, 03 = no spouse or unmarried partner present, 99 = NIU). We exclude observations coded as NIU (SPOUSEPRES = 99) and use "no spouse or unmarried partner present" (SPOUSEPRES = 03) as the reference category. We then construct two binary indicator variables, *Spouse present* and *Unmarried partner present*, which take the value 1 when SPOUSEPRES equals 01 or 02, respectively, and 0 otherwise. These indicators enter the models as covariates capturing household partnership status.

Feature scaling and coding. Before estimation, we standardize and encode all covariates in a consistent way. On the continuous side, we treat respondents' age, educational attainment, and weekly earnings as continuous predictors and include age, the four-level education variable edu, and weekly earnings EARNWEEK. These three variables are standardized to have mean zero and unit variance using the sample mean and standard deviation from the final analytic sample. This scaling improves numerical stability and makes the corresponding coefficients more comparable in magnitude across covariates.

Binary and categorical variables are kept on an interpretable 0/1 scale and are not standardized. Sex enters the models as a binary indicator *Male*, coded as 1 for respondents coded as male in SEX and 0 for those coded as female (the reference category). Household partnership status, originally recorded in SPOUSEPRES, is represented by two indicator variables, *Spouse present* and *Unmarried partner present*, with "no spouse or unmarried partner present" serving as the baseline category. Race, originally coded in RACE, is represented by four single-race indicators—*Race: Black*, *Race: Native American*, *Race: Asian*, and *Race: Pacific Islander*—with "White only" as the omitted reference group. All standardized continuous variables and dummy-coded indicators are concatenated into a single design matrix, and we include an explicit intercept term for use in the subsequent empirical analysis and LLM-based simulations.

D Descriptive Statistics

Table 7 reports descriptive statistics for respondent attributes and time-use outcomes in the final analytic sample of 4,307 individuals. By construction, the continuous attributes (age, education, and weekly earnings) are standardized to have mean zero and unit variance. The binary attributes indicate that roughly half of the sample is male (mean ≈ 0.51), just over half live with a spouse (≈ 0.54), and a small fraction live with an unmarried partner (≈ 0.06). The racial composition is predominantly White, with comparatively small shares of Black, Asian, Native American, and Pacific Islander respondents. On average, respondents allocate about 247 minutes per day to work, 266 minutes to leisure, 578 minutes to sleep and personal care, and 349 minutes to other activities, with substantial cross-individual variation in both work and non-work time.

E Additional Details of Experimental Setup

LLMs under evaluation. We evaluate a diverse set of LLMs to demonstrate that XCHOICE applies across major model families and capa-

Table 8: Estimated Coefficients by Activity and Model (Human Baseline vs. LLMs)

Feature	Human	GPT-4o	Claude-3.7	Deepseek-V3	Llama-3.3	Qwen-2.5
Activity: Leisure						
Constant	**0.326** ± 0.102	**1.035** ± 0.041	**1.344** ± 0.043	**1.704** ± 0.090	**1.331** ± 0.088	**1.752** ± 0.051
Age	**0.055** ± 0.029	**-0.146** ± 0.011	**0.102** ± 0.012	**0.050** ± 0.024	**0.066** ± 0.023	**0.038** ± 0.014
Sex	**-0.324** ± 0.059	**-0.157** ± 0.023	**-0.225** ± 0.024	**-0.186** ± 0.050	**-0.168** ± 0.049	**-0.141** ± 0.028
Education	-0.025 ± 0.032	**-0.014** ± 0.012	**-0.044** ± 0.013	**0.044** ± 0.027	**0.059** ± 0.026	0.012 ± 0.016
Earnings	**-0.060** ± 0.033	**0.069** ± 0.013	**0.045** ± 0.014	-0.002 ± 0.029	**0.040** ± 0.029	**-0.041** ± 0.016
Asian	-0.072 ± 0.118	**-0.132** ± 0.048	**-0.103** ± 0.048	-0.075 ± 0.100	**0.125** ± 0.122	-0.052 ± 0.055
Black	**0.164** ± 0.098	**-0.053** ± 0.036	**-0.100** ± 0.039	**-0.119** ± 0.077	-0.022 ± 0.077	**-0.051** ± 0.045
Native Am.	0.027 ± 0.321	0.088 ± 0.139	-0.038 ± 0.138	-0.154 ± 0.261	**-0.299** ± 0.216	**-0.574** ± 0.108
Pac. Isl.	0.130 ± 0.795	0.026 ± 0.282	-0.005 ± 0.297	-0.067 ± 0.517	-0.164 ± 0.470	**-0.390** ± 0.265
Married	**-0.249** ± 0.062	**-0.180** ± 0.023	**-0.210** ± 0.025	**-0.106** ± 0.052	-0.017 ± 0.049	**-0.063** ± 0.030
Unmarried	0.036 ± 0.131	**-0.141** ± 0.055	**-0.122** ± 0.054	0.036 ± 0.113	-0.030 ± 0.105	-0.051 ± 0.061
Activity: Sleep and Personal Care						
Constant	**0.677** ± 0.074	**1.537** ± 0.036	**1.942** ± 0.039	**2.273** ± 0.085	**1.112** ± 0.090	**1.912** ± 0.050
Age	-0.011 ± 0.021	**-0.226** ± 0.010	**-0.014** ± 0.011	-0.009 ± 0.023	0.005 ± 0.023	**0.019** ± 0.014
Sex	**-0.064** ± 0.042	**-0.089** ± 0.020	**-0.130** ± 0.022	**-0.127** ± 0.048	**-0.076** ± 0.050	**-0.116** ± 0.028
Education	**-0.038** ± 0.023	**-0.021** ± 0.011	**-0.053** ± 0.012	**0.049** ± 0.026	**0.037** ± 0.027	0.006 ± 0.015
Earnings	-0.018 ± 0.023	**0.108** ± 0.012	**0.091** ± 0.013	0.008 ± 0.027	0.028 ± 0.030	**-0.059** ± 0.016
Asian	0.003 ± 0.078	**-0.052** ± 0.041	-0.019 ± 0.042	-0.036 ± 0.094	0.081 ± 0.124	-0.021 ± 0.054
Black	**0.189** ± 0.072	**0.040** ± 0.031	0.001 ± 0.034	-0.039 ± 0.072	0.019 ± 0.077	-0.039 ± 0.044
Native Am.	0.033 ± 0.228	0.089 ± 0.123	0.031 ± 0.123	-0.082 ± 0.243	**-0.247** ± 0.217	**-0.531** ± 0.103
Pac. Isl.	0.433 ± 0.559	0.052 ± 0.251	0.074 ± 0.266	-0.031 ± 0.486	-0.081 ± 0.468	-0.219 ± 0.252
Married	**-0.187** ± 0.044	**0.049** ± 0.021	**-0.105** ± 0.023	-0.023 ± 0.049	-0.005 ± 0.050	**-0.035** ± 0.029
Unmarried	0.008 ± 0.096	0.006 ± 0.048	**-0.071** ± 0.048	-0.014 ± 0.107	-0.041 ± 0.106	-0.051 ± 0.060
Activity: Work						
Constant	0.076 ± 0.108	**1.265** ± 0.038	**1.741** ± 0.040	**2.127** ± 0.086	**2.152** ± 0.082	**2.043** ± 0.049
Age	0.009 ± 0.031	**-0.353** ± 0.010	**-0.104** ± 0.011	**-0.115** ± 0.023	**-0.169** ± 0.021	**-0.092** ± 0.014
Sex	**-0.221** ± 0.063	**-0.109** ± 0.021	**-0.164** ± 0.022	**-0.146** ± 0.048	**-0.384** ± 0.046	**-0.116** ± 0.028
Education	**-0.099** ± 0.034	**-0.029** ± 0.011	**-0.081** ± 0.012	**0.033** ± 0.026	**0.072** ± 0.025	-0.011 ± 0.015
Earnings	**0.112** ± 0.034	**0.184** ± 0.012	**0.175** ± 0.013	**0.068** ± 0.028	**0.234** ± 0.028	**0.037** ± 0.015
Asian	-0.045 ± 0.120	**-0.074** ± 0.042	0.016 ± 0.043	-0.013 ± 0.094	**0.371** ± 0.113	-0.038 ± 0.053
Black	**0.216** ± 0.101	**0.059** ± 0.033	**0.075** ± 0.035	-0.008 ± 0.073	**0.163** ± 0.071	0.002 ± 0.043
Native Am.	0.055 ± 0.336	**0.131** ± 0.127	0.095 ± 0.125	-0.051 ± 0.245	**-0.331** ± 0.197	**-0.436** ± 0.100
Pac. Isl.	0.319 ± 0.769	0.139 ± 0.257	0.120 ± 0.271	-0.213 ± 0.497	-0.185 ± 0.434	**-0.256** ± 0.250
Married	**-0.236** ± 0.065	**0.046** ± 0.022	**-0.119** ± 0.023	-0.030 ± 0.050	0.037 ± 0.046	**-0.036** ± 0.029
Unmarried	0.031 ± 0.135	-0.011 ± 0.050	**-0.095** ± 0.049	0.006 ± 0.108	-0.039 ± 0.096	-0.056 ± 0.059

Note: Coefficients significant at $p < 0.05$ are **bolded**. Margins denote 95% CI.

bility levels. For closed-source systems, we include `gpt-4o-2024-08-06` (*GPT-4o*) and `claude-3-7-sonnet-20250219` (*Claude-3.7*). We also assess leading open source models, including `Qwen/Qwen2.5-72B-Instruct` (*Qwen-2.5*), `meta-llama/Llama-3.3-70B-Instruct` (*Llama-3.3*), and `deepseek-ai/DeepSeek-V3` (*DeepSeek-V3*). When using provider-specific endpoints, we report the underlying base model name for clarity and reproducibility.

Generation and filtering protocol. For each ATUS respondent profile, we issue a single query to each LLM using the prompt template in Appendix A. Decoding parameters are held constant across all five main-experiment models: temperature 0.1, top-p 1.0, and a maximum of 1,024 tokens, matching the configuration used in the RAG intervention in Appendix H. We do not fix a global random seed across model calls; instead, robustness to sampling variability is addressed by the non-parametric bootstrap analysis in Appendix I.5, which directly quantifies uncertainty in the recovered parameters and diagnostic rankings.

The prompt constrains the response to the structured format `[Work, Leisure, Sleep and Personal Care, Other]` in minutes. Across all $N = 4{,}307$ profiles and all five evaluated LLMs, every returned response parses successfully and strictly satisfies the time-budget constraint with non-negative integer allocations summing to exactly 1,440 minutes. No retries, repairs, or filtering are required, yielding a retention rate of 100%. Zero entries do occur for some respondents and activities under both the human ATUS data and the

LLM outputs; because such corner solutions interact with the log specification in Eq. 4, we examine their effect on the recovered parameters and diagnostic rankings via a small smoothing perturbation in Appendix I.3.

F Additional Experimental Results

Full coefficient estimates. Table 8 reports the full coefficient estimates for each activity-specific model, including the human baseline and all evaluated LLMs. Coefficients are presented as point estimates with 95% confidence intervals, and statistically significant estimates ($p < 0.05$) are boldfaced. We include this table as supplementary evidence for the attribute-level comparisons in the main text (Figure 3 and Table 3), enabling readers to inspect the exact magnitudes, signs, and uncertainty of each demographic factor across activities and models.

Formal tests of human–LLM coefficient differences. To test the headline subgroup claims directly, we run Wald tests of $H_0 : \theta_{l,f} = \theta_{H,f}$ for every model l, activity, and the two focal features *Race: Black* and *Spouse: Present* (30 tests), using $z = (\hat{\theta}_{l,f} - \hat{\theta}_{H,f})/\sqrt{\mathrm{SE}_l^2 + \mathrm{SE}_H^2}$ with standard errors recovered from the 95% confidence-interval margins in Table 8 ($\mathrm{SE} = \mathrm{margin}/1.96$). We apply the Holm correction across all 30 tests. Table 10 reports the results: 27 of 30 differences remain significant after correction, 16 of them at adjusted $p < 0.001$. The three non-significant cells (GPT-4o and Claude-3.7 on *Spouse: Present* $\times$ *Leisure*, and Llama-3.3 on *Race: Black* $\times$ *Work*) are all cells that Table 4 does not mark as worst-aligned, while every worst-aligned cell is significant, so the formal tests and the descriptive diagnostics agree.

Decision-weight profiles by model. Figure 5 complements Figure 3 by reorganizing the same coefficient estimates at the *model* level. Each panel corresponds to one decision maker (Human or an LLM), and the x-axis lists features, while colors indicate activities (Work, Leisure, and Sleep/Personal Care) with 95% confidence intervals. Unlike the main-text attribute-level plot, we include *Race: Pacific Islander* and *Race: Native American* here for completeness; the wide intervals and unstable estimates illustrate why these categories are omitted from the main results due to limited sample size and lack of meaningful statistical precision. Overall, the model-level view makes it easier to inspect how each model reweights attributes across activities and where uncertainty or sign changes concentrate.

Distributional view of misalignment identification. Figures 6 and Figure 7 provide a distributional view of the misalignment identified by the coefficient-based diagnostics. Across both race and partnership-status strata, human time shares (black curves) are noticeably more diffuse, with heavier tails that reflect heterogeneous schedules and constraints. In contrast, LLM-generated shares are highly concentrated, showing clear heaping at a few focal values, which is consistent with stereotyped allocations such as round-hour schedules. A prominent pattern is that models systematically under-allocate the residual category *Other*: compared with the human distributions, LLM densities collapse toward near-zero time shares and rarely reproduce the wider support and long right tail observed in the survey data. This suggests that models tend to over-commit time to the salient categories (Work/Leisure/Personal) while failing to capture the diversity of non-core activities that occupy substantial portions of the day for many individuals. The subgroup panels also highlight why *Race: Black* and *Spouse: Present* emerge as salient divergence sources in the mechanism estimates. In the human data, the black curves shift in shape across strata, indicating that demographic attributes are associated with meaningful changes in how time is distributed across activities (including tail behavior and multimodality). By comparison, many LLM curves vary only weakly across rows, implying limited sensitivity to subgroup-specific trade-offs even when demographic information is provided. Put differently, the models often produce similar "template" schedules across groups, compressing within-group variation and attenuating cross-group differences. This distributional evidence is consistent with the coefficient misalignment results: the models do not merely miss outcomes at the margin, but encode different prioritization patterns for these subgroups, especially in how they allocate residual time and how strongly demographic attributes reshape the allocation trade-offs.

OLS estimation results. Table 9 reports reduced-form OLS estimates for each activity and decision maker, including the human baseline and LLM-generated decisions. For each activity, we regress the observed or model-produced time share on the same set of demographic covariates (standardized

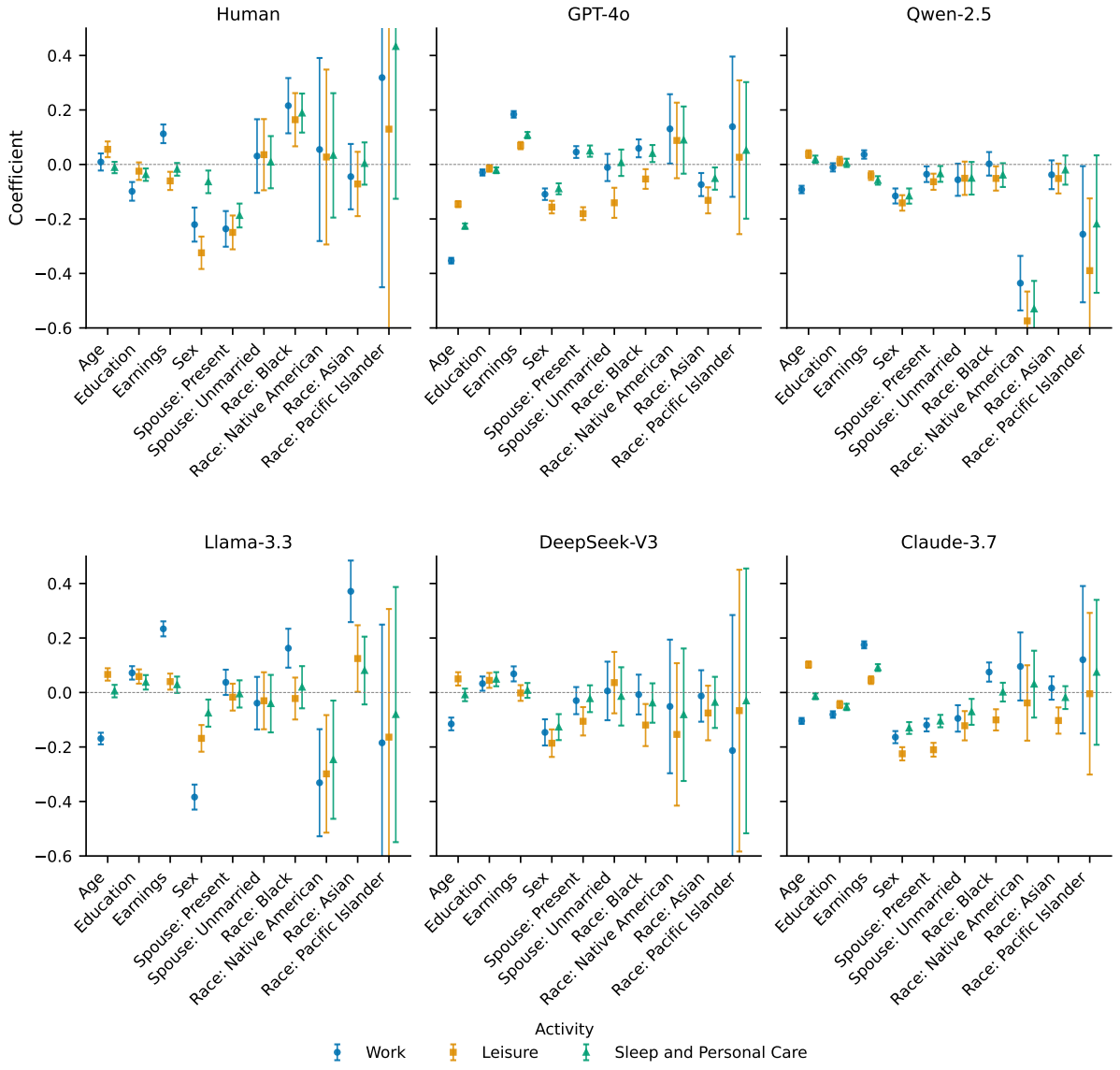


Figure 5: Estimated Decision-Weight Profiles by Model (Across Activities and Attributes)

age, education, weekly earnings, plus indicators for sex, spouse or partner status, and race), and report coefficients with 95% confidence-interval margins. These coefficients summarize conditional associations and do not enforce the time-budget adding-up constraint across activities, so they should not be interpreted as trade-off weights. Instead, we use them as a conventional baseline to contrast against XCHOICE's mechanism-based estimates and to support robustness checks under mild environment shifts. The table also highlights that coefficients for sparsely represented groups, especially Native American and Pacific Islander, can be highly variable and imprecise, which aligns with our decision to omit these categories from the main-text attribute-level visualization.

G Invariance Analysis Metrics and Experimental Design

We test parameter invariance by comparing baseline estimates $\theta^{(0)}$ to estimates $\theta^{(1)}$ obtained after applying mild counterfactual data shifts. Drift is quantified by three lower-is-better metrics:

$$\mathrm{MAD} = \frac{1}{K}\sum_{i=1}^{K}\left|\theta_i^{(1)} - \theta_i^{(0)}\right|, \tag{6}$$

$$\mathrm{RelL2} = \frac{\left\|\theta^{(1)} - \theta^{(0)}\right\|_2}{\left\|\theta^{(0)}\right\|_2}, \tag{7}$$

$$1 - \mathrm{CosSim} = 1 - \frac{\theta^{(1)} \cdot \theta^{(0)}}{\left\|\theta^{(1)}\right\|_2 \left\|\theta^{(0)}\right\|_2}. \tag{8}$$

MAD measures average absolute coefficient

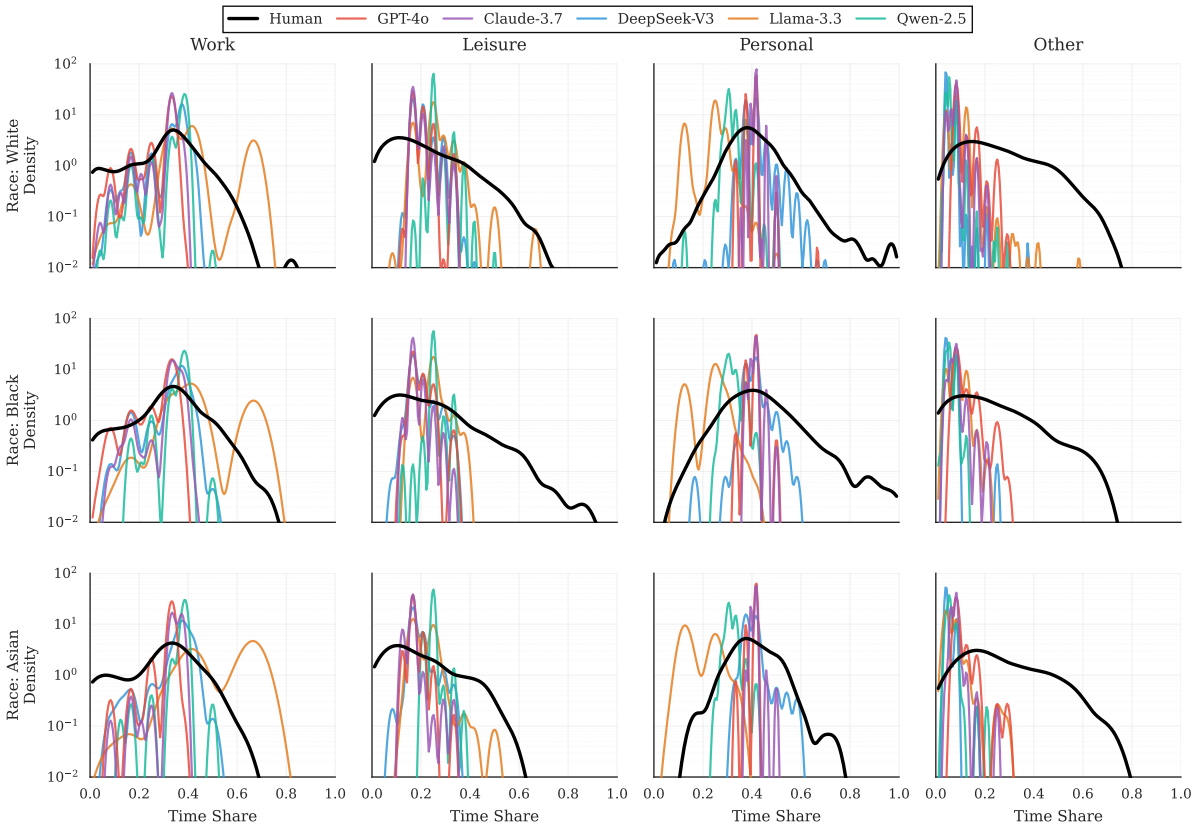


Figure 6: Time-share distributions across activities by race group. Each column corresponds to one activity (Work, Leisure, Personal, Other), and each row to a race group. Within each panel, we compare the empirical distribution of human time shares (black) with time-share patterns produced by different LLM-based models, using a log-scaled density axis.

change, RelL2 measures relative change in ℓ_2 magnitude, and $1 - \mathrm{CosSim}$ measures directional rotation (near 0 implies preserved relative trade-offs).

Data shift design. We implement four shifts that emulate small policy-like perturbations and compositional changes:

- **Earnings quantile lift (magnitude 0.5).** Let $q_{0.5}$ be the median of Weekly earnings (`earnweek`). We only lift the bottom half by adding $0.5 \cdot \mathrm{sd}(\texttt{earnweek})$ to all individuals with $\texttt{earnweek} \leq q_{0.5}$, leaving the upper half unchanged. This changes the *shape* of the earnings distribution (not a pure location shift).
- **Age band shift (magnitude 0.1).** We randomly select a fraction $p = 0.1$ of individuals aged 25–34 and increase their continuous age by 10 years, effectively moving them into the 35–44 range. This produces a localized age redistribution and can modify correlations between age and other covariates (e.g., spouse/education/earnings).
- **Race mix shift (Asian +2pp, White −2pp).** After discretizing age into `age_band`, we perform a mild raking-style resampling within each (`age_band`, `sex`) stratum to increase race code 131 (Asian) by 0.02 in share and decrease race code 100 (White; baseline) by 0.02, keeping other categories unchanged up to normalization. This is a small compositional (mixture) shift that preserves the stratification structure.
- **Spouse mix shift (present +3pp in ages 30–44).** Using `age_band` bins $(18, 25, 30, 35, 45, 55, 65, 100)$, we apply raking-style resampling *only* within age bands covering 30–44 to increase $\texttt{spousepres} = 1$ (present) by 0.03 and decrease $\texttt{spousepres} = 3$ (baseline) by 0.03. Other age bands are left unchanged. This creates a targeted, policy-like compositional perturbation rather than a global reweighting.

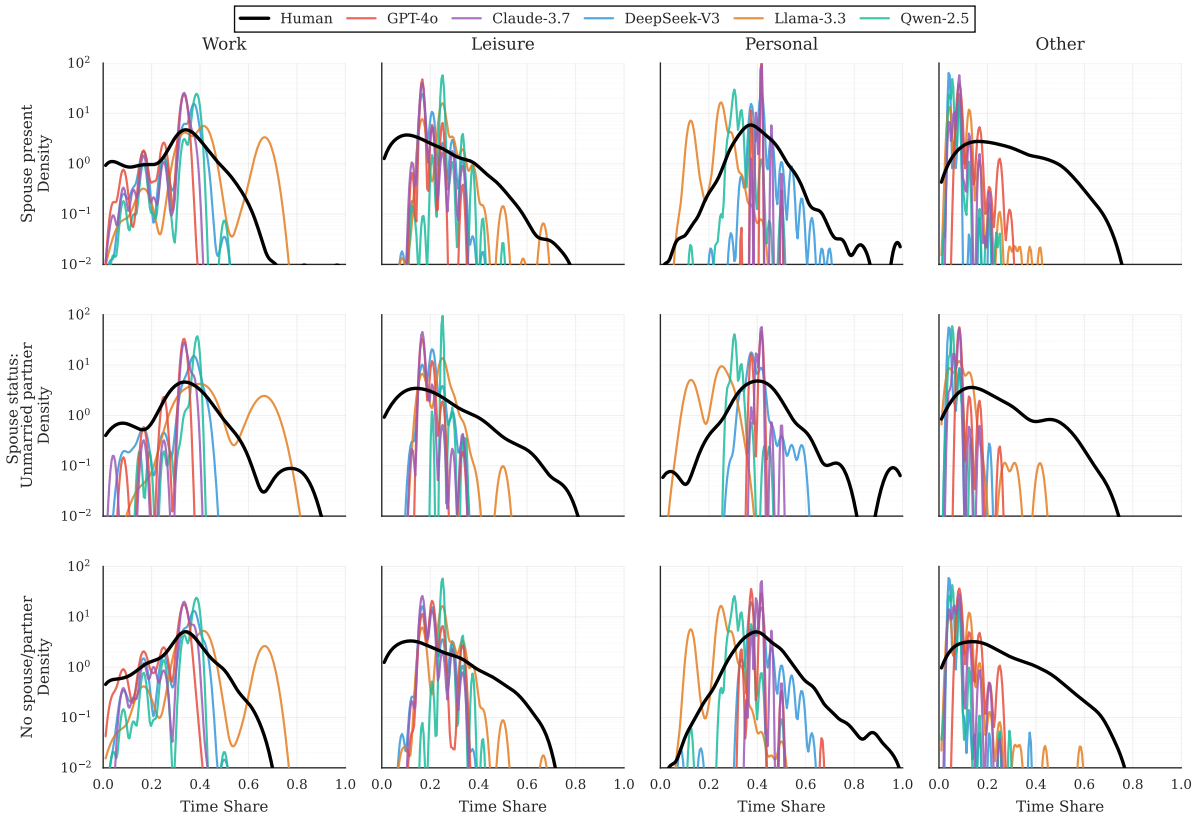


Figure 7: Time-share distributions across activities by household partnership status. Each column corresponds to one activity (Work, Leisure, Personal, Other), and each row to a spouse/partner status category (spouse present, unmarried partner, no spouse/partner). Within each panel, we compare the empirical distribution of human time shares (black) with time-share patterns produced by different LLM-based models, using a log-scaled density axis.

H RAG Design for Misalignment Mitigation

To analyze alignment at the attribute (and subgroup) level, we compare how each model weights the same feature across activities. For a given feature f (e.g., *Race: Black* or *Spouse: Present*), we form the activity-specific coefficient vector $\hat{\theta}_{m,f} = (\hat{\theta}^L_{m,f}, \hat{\theta}^W_{m,f}, \hat{\theta}^S_{m,f})$ for each model m, and likewise $\hat{\theta}_{H,f}$ for the human baseline. We then quantify alignment using cosine similarity:

$$CosSim(\hat{\theta}_{H,f}, \hat{\theta}_{m,f}) = \frac{\hat{\theta}^\top_{H,f}\hat{\theta}_{m,f}}{\|\hat{\theta}_{H,f}\| \, \|\hat{\theta}_{m,f}\|}.$$

This metric captures whether humans and LLMs exhibit similar *directional* trade-off patterns for feature f across activities, independent of overall scale, and enables systematic diagnosis of which attributes drive mechanism-level divergence.

We employ a Retrieval-Augmented Generation (RAG) pipeline to enhance LLM predictions of time allocation by grounding model outputs in empirical research findings. The following subsections describe the construction of our knowledge base, the retrieval mechanism, and the prompt design.

H.1 Knowledge Base Construction

We construct domain-specific knowledge bases from peer-reviewed research on time allocation behavior. Each knowledge base contains structured instances extracted from empirical studies, capturing findings about how demographic factors influence daily time use patterns.

Marriage Knowledge Base. This knowledge base comprises 43 instances derived from research papers examining the relationship between marital status and time allocation. Topics include household decision-making dynamics, work hours by marital status, couple time and family activities, the impact of children on time allocation, leisure time patterns, personal care behaviors, and caregiving responsibilities.

Race Knowledge Base. This knowledge base contains 50 instances from studies investigating

Table 9: Estimated Coefficients by Activity and Model (Human Baseline vs. LLMs). Values represent OLS coefficients ± 95% CI margin.

Feature	Human	GPT-4o	Claude	DeepSeek	Llama	Qwen
Activity: Leisure						
Age	**0.061** ± 0.028	**0.115** ± 0.006	**0.151** ± 0.006	**0.110** ± 0.008	**0.190** ± 0.010	**0.096** ± 0.005
Sex	**-3.508** ± 0.841	**-0.894** ± 0.185	**-1.280** ± 0.164	**-0.805** ± 0.248	**1.483** ± 0.289	**-0.458** ± 0.135
Education	0.171 ± 0.425	**0.131** ± 0.094	**0.164** ± 0.083	0.048 ± 0.125	0.088 ± 0.146	**0.255** ± 0.068
Earnings	**-0.001** ± 0.001	**-0.001** ± 0.000	**-0.001** ± 0.000	**-0.001** ± 0.000	**-0.002** ± 0.000	**-0.001** ± 0.000
Asian	-0.950 ± 1.641	**-1.079** ± 0.362	**-1.376** ± 0.321	**-0.783** ± 0.483	**-2.420** ± 0.565	**-0.465** ± 0.264
Black	0.465 ± 1.335	**-1.509** ± 0.294	**-1.855** ± 0.261	**-1.566** ± 0.393	**-2.150** ± 0.459	**-0.658** ± 0.214
Native Am.	0.135 ± 4.581	-0.134 ± 1.009	**-1.244** ± 0.896	**-1.355** ± 1.348	-0.443 ± 1.575	**-2.646** ± 0.734
Pac. Isl.	-2.182 ± 9.411	-0.694 ± 2.073	-1.283 ± 1.840	0.400 ± 2.769	-0.684 ± 3.236	**-3.225** ± 1.507
Married	**-1.600** ± 0.879	**-3.515** ± 0.194	**-1.669** ± 0.172	**-1.327** ± 0.259	**-0.793** ± 0.302	**-0.500** ± 0.141
Unmarried	0.606 ± 1.818	**-2.175** ± 0.401	**-0.712** ± 0.356	**0.685** ± 0.537	0.042 ± 0.626	0.062 ± 0.292
Activity: Sleep and Personal Care						
Age	**-0.049** ± 0.019	**0.010** ± 0.004	**0.025** ± 0.003	**0.067** ± 0.008	**0.101** ± 0.011	**0.079** ± 0.005
Sex	**2.001** ± 0.575	**0.599** ± 0.112	**0.471** ± 0.084	**0.652** ± 0.235	**3.415** ± 0.319	0.138 ± 0.163
Education	-0.023 ± 0.291	-0.008 ± 0.057	**0.166** ± 0.042	**0.301** ± 0.119	**-0.364** ± 0.162	**0.139** ± 0.082
Earnings	**-0.001** ± 0.000	**-0.000** ± 0.000	**-0.001** ± 0.000	**-0.001** ± 0.000	**-0.003** ± 0.000	**-0.002** ± 0.000
Asian	0.863 ± 1.124	**0.654** ± 0.219	-0.034 ± 0.163	-0.100 ± 0.458	**-3.022** ± 0.624	**0.419** ± 0.319
Black	**1.604** ± 0.914	**0.681** ± 0.178	**-0.336** ± 0.133	0.144 ± 0.373	**-1.147** ± 0.507	**-0.457** ± 0.258
Native Am.	0.244 ± 3.136	-0.178 ± 0.611	-0.426 ± 0.456	-0.065 ± 1.277	0.738 ± 1.740	**-1.945** ± 0.887
Pac. Isl.	**7.311** ± 6.442	-0.536 ± 1.255	0.047 ± 0.936	2.563 ± 2.624	1.467 ± 3.574	1.180 ± 1.822
Married	**-1.289** ± 0.602	**2.234** ± 0.117	**0.615** ± 0.087	**0.828** ± 0.245	-0.282 ± 0.334	**0.281** ± 0.170
Unmarried	-0.563 ± 1.244	**1.544** ± 0.243	**0.435** ± 0.181	**-0.621** ± 0.509	-0.062 ± 0.692	0.055 ± 0.353
Activity: Work						
Age	0.000 ± 0.038	**-0.272** ± 0.011	**-0.191** ± 0.009	**-0.200** ± 0.014	**-0.329** ± 0.023	**-0.194** ± 0.008
Sex	**-1.996** ± 1.115	-0.147 ± 0.316	**-0.591** ± 0.278	-0.166 ± 0.402	**-6.728** ± 0.683	0.127 ± 0.246
Education	**-0.939** ± 0.564	**-0.288** ± 0.160	**-0.769** ± 0.141	**-0.304** ± 0.203	**0.622** ± 0.346	**-0.438** ± 0.125
Earnings	**0.002** ± 0.001	**0.003** ± 0.000	**0.003** ± 0.000	**0.002** ± 0.000	**0.006** ± 0.000	**0.002** ± 0.000
Asian	-0.430 ± 2.177	-0.144 ± 0.618	**1.144** ± 0.543	0.768 ± 0.784	**6.669** ± 1.335	-0.156 ± 0.482
Black	1.285 ± 1.770	**1.145** ± 0.502	**2.191** ± 0.442	**1.184** ± 0.638	**3.969** ± 1.084	**1.051** ± 0.391
Native Am.	0.387 ± 6.076	1.333 ± 1.724	**1.896** ± 1.516	1.088 ± 2.187	-2.419 ± 3.720	1.265 ± 1.341
Pac. Isl.	0.798 ± 12.482	2.287 ± 3.541	1.595 ± 3.114	-4.069 ± 4.493	-1.736 ± 7.642	0.058 ± 2.755
Married	**-1.375** ± 1.166	**1.525** ± 0.331	0.020 ± 0.291	0.359 ± 0.420	**1.001** ± 0.714	0.197 ± 0.257
Unmarried	0.117 ± 2.411	0.563 ± 0.684	-0.518 ± 0.602	0.065 ± 0.871	-0.293 ± 1.479	-0.226 ± 0.533

racial and ethnic disparities in time use. Topics include leisure-time physical activity disparities, work-related activity patterns, time penalties faced by racial minorities, employment barriers, structural factors affecting time allocation, gender differences in household activities, and sleep duration variations across racial groups.

H.2 Embedding and Retrieval

We employ dense retrieval to identify relevant knowledge instances for each respondent persona.

Embedding Generation. We generate embeddings for both knowledge instances and respondent personas using OpenAI's `text-embedding-3-large` model, which produces 3,072-dimensional dense vector representations. Each knowledge instance is embedded based on its textual content, while respondent personas are embedded using a structured sentence template:

> *"A [age]-year-old [gender] [race] with [education], [spouse status], earning $[income] per week."*

Retrieval Mechanism. For each respondent, we compute cosine similarity between the persona embedding and all knowledge instance embeddings. We retrieve the top-k most similar instances, where $k = 3$ in our experiments.

H.3 Demographic Variable Encoding

We encode demographic variables from the ATUS dataset into natural language descriptions. Gender is encoded as male or female. Education level is categorized into four groups: no college education, some college without a bachelor's degree, bachelor's degree, and advanced degree. Race categories include White, Black, American Indian or Alaskan

Table 10: Wald statistics z for human–LLM coefficient differences on the two focal features. Stars denote Holm-adjusted significance across all 30 tests: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Sleep/PC denotes *Sleep and Personal Care*.

Activity	Feature	GPT-4o	Claude-3.7	DeepSeek-V3	Llama-3.3	Qwen-2.5
Leisure	Race: Black	-4.07^{***}	-4.91^{***}	-4.45^{***}	-2.93^{*}	-3.91^{**}
Leisure	Spouse: Present	2.05	1.14	3.46^{**}	5.75^{***}	5.29^{***}
Sleep/PC	Race: Black	-3.73^{**}	-4.63^{***}	-4.39^{***}	-3.16^{*}	-5.30^{***}
Sleep/PC	Spouse: Present	9.49^{***}	3.24^{**}	4.88^{***}	5.36^{***}	5.65^{***}
Work	Race: Black	-2.90^{*}	-2.59^{*}	-3.52^{**}	-0.84	-3.82^{**}
Work	Spouse: Present	8.05^{***}	3.33^{**}	4.92^{***}	6.72^{***}	5.51^{***}

Native, Asian, and Native Hawaiian or Pacific Islander. Spouse presence is encoded as having a spouse, having an unmarried partner, or having no spouse or unmarried partner.

H.4 Model Configuration

We evaluate two LLMs in our experiments: GPT-4o and DeepSeek-V3. Both models are configured with a temperature of 0.1 to promote deterministic outputs, a maximum token limit of 1,024, and top-p sampling set to 1. These settings balance output consistency with sufficient generation capacity for the structured JSON response.

H.5 Illustrative Examples

We present examples demonstrating how the RAG pipeline retrieves contextually relevant knowledge for different demographic profiles. For each example, we show the respondent's demographic information, the constructed persona sentence used for embedding, and the full text of the three retrieved knowledge instances.

H.5.1 Marriage Experiment Examples

Example 1: Married Individual. Demographics: 53-year-old Black male with an advanced degree, married, weekly income of $1,173.
Persona sentence: *"A 53-year-old male Black with an advanced degree, a spouse, earning $1173.00 per week."*
Retrieved knowledge instances:

1. *"Expected wage profiles at the time of marriage strongly impact household decision-making weights, favoring the spouse with a higher expected wage."*

2. *"Dual-earner households exhibit more frequent adjustments in weights due to wage shocks compared to single-earner households."*

3. *"Understanding how marital status and wage profiles affect intra-household decision-making can inform interventions designed to improve family welfare and resource allocation."*

Example 2: Unmarried Individual. Demographics: 39-year-old White female with some college education but no bachelor's degree, no spouse or unmarried partner, weekly income of $240.
Persona sentence: *"A 39-year-old female White with some college education but no bachelor's degree, no spouse or unmarried partner, earning $240.00 per week."*
Retrieved knowledge instances:

1. *"Married individuals, particularly husbands, tend to work longer hours compared to their unmarried peers."*

2. *"Married individuals often have more stable employment patterns, prioritizing job security to support their families. Unmarried individuals tend to engage in more part-time work or gig economy jobs, reflecting a more flexible approach to employment."*

3. *"Leisure hours are relatively similar between genders, with slight differences in how they allocate time to leisure activities."*

H.5.2 Race Experiment Examples

Example 3: Black Individual. Demographics: 53-year-old Black male with an advanced degree, married, weekly income of $1,173.
Persona sentence: *"A 53-year-old male Black with an advanced degree, a spouse, earning $1173.00 per week."*
Retrieved knowledge instances:

1. *"Leisure activities account for 53.8% of the time for urban African American adolescents, which is higher than in other postin-*

dustrial populations. Significant engagement in leisure activities includes TV viewing (17.5%) and talking (11.1%). Boys reported more leisure time than girls among the urban African American adolescents."

2. *"Urban African American adolescents spent 21.6% of their time on productive activities, with only 4.5% dedicated to homework, which is less than their suburban peers."*

3. *"Urban African American adolescents spent 3.99% of their time on household chores and 7.89% on personal maintenance, with girls reporting more time in these activities."*

Example 4: Non-Black Individual. Demographics: 30-year-old White female with a bachelor's degree, unmarried partner, weekly income of $1,730.76.
Persona sentence: *"A 30-year-old female White with a bachelor's degree, an unmarried partner, earning $1730.76 per week."*
Retrieved knowledge instances:

1. *"Women, particularly Spanish-speaking Hispanics, report higher engagement in heavy household chores compared to their White counterparts, suggesting a gendered aspect of time allocation in household responsibilities."*

2. *"Non-Whites report higher levels of work-related physical activity (WRPA), with Spanish-speaking Hispanics engaging in strenuous job-related activities most frequently. Whites report the lowest levels of WRPA compared to other racial and ethnic groups."*

3. *"Urban African American adolescents spent 3.99% of their time on household chores and 7.89% on personal maintenance, with girls reporting more time in these activities."*

H.6 Paired Estimation Protocol

RAG generation is run on a random 10% subsample of the analytic sample ($N = 431$). To separate the effect of the intervention from sampling differences, baseline and post-intervention parameters are estimated on the same respondents: baseline parameters use the main-experiment outputs of the corresponding model for these 431 respondents, post-intervention parameters use the RAG outputs, and both use the covariate coding of the main pipeline. The human reference is the full-sample estimate. Cosine similarities are computed on the focal-feature coefficient vector defined above, and 95% percentile intervals are obtained from $B = 200$ bootstrap replicates that resample respondents jointly across arms.

H.7 Qualitative-Evidence Ablation

To test whether post-RAG shifts reflect imitation of quoted statistics rather than adjustment of trade-off weights, we construct a qualitative arm. Retrieval is held identical: each respondent receives the same three retrieved passages as in the original run. Within these passages, every quantitative expression (e.g., percentages, hours, and minutes) is paraphrased into a qualitative statement conveying the same directional finding, and passages without quantitative content are left unchanged. The prompt template, models, and decoding configuration are identical across arms, so the passage text is the only difference. We then estimate parameters on the qualitative-arm outputs under the same paired protocol and bootstrap the between-arm difference jointly with the other quantities.
Table 6 reports the results. In the one cell with a credible gain (GPT-4o on *Spouse: Present*), the qualitative arm retains 96% of the gain, and the between-arm difference is indistinguishable from zero (95% CI $[-0.13, 0.11]$), so the gain does not depend on quoted statistics. In DeepSeek-V3 on *Spouse: Present*, removing statistics eliminates most of the degradation (between-arm difference 95% CI $[0.19, 1.34]$), indicating that the statistics themselves distort an already human-consistent mechanism. In GPT-4o on *Race: Black*, the degradation persists in both arms (between-arm difference 95% CI $[-0.03, 0.15]$), consistent with a failure driven by the content of mismatched evidence rather than its format.

Output validity. Retrieval-augmented prompts lower the rate of usable outputs relative to the main experiment. Table 11 reports, for each condition, the number of respondents whose output yields a parseable allocation with a positive total; shares are computed by normalizing each allocation by its total. Post-intervention estimates use these respondents, and baseline estimates use all 431 respondents.

Table 11: Usable post-intervention outputs out of $N = 431$ respondents per condition.

Model	Subgroup	RAG	RAG (qual.)
GPT-4o	Spouse: Present	424 (98.4%)	410 (95.1%)
GPT-4o	Race: Black	344 (79.8%)	339 (78.7%)
DeepSeek-V3	Spouse: Present	393 (91.2%)	376 (87.2%)
DeepSeek-V3	Race: Black	379 (87.9%)	379 (87.9%)

I Robustness Checks and Additional Validations

This appendix consolidates a series of robustness checks that probe whether the diagnostic conclusions in Section 4 depend on specific modeling or identification choices. Throughout, we hold the LLM estimates fixed and re-estimate the human mechanism under alternative configurations, then recompute the same feature-wise multi-model averaged deviation

$$A_f = \frac{1}{L \cdot J^\dagger} \sum_{l,j} \left| \hat{\theta}_{l,f,j} - \hat{\theta}_{H,f,j} \right|,$$

where L is the number of LLMs and $J^\dagger$ is the set of non-reference activities. Robustness is summarized by Top-5 overlap, Jaccard, Spearman rank correlation across all features, and the maximum rank shift within the Top-K union, following Webber et al. (2010).

I.1 Identification Sensitivity: Baseline Swap

The structural parameters in Eq. 4 are identified up to a reference-category normalization. The main text fixes $\theta_O = 0$ for *Other Activities*. To verify that the misalignment ranking is not an artifact of this specific choice, we re-estimate all parameters under the equivalent normalization $\theta_S = 0$ that uses *Sleep and Personal Care* as the reference. We then recompute A_f over the same set of demographic features.

Table 12 reports the four robustness indicators. The Top-5 set of most misaligned features is identical across the two normalizations, with a Jaccard index of 1.00. The Spearman rank correlation across all ten features is 0.879 and the Kendall tau is 0.733, indicating strong concordance. The maximum rank shift on the Top-K union is 3 positions, reflecting mild local reordering rather than substantive instability. Table 13 provides the full Top-10 side-by-side comparison: the same ten features appear in both rankings, and the only reordering occurs among adjacent positions inside two near-tie clusters. We note that the absolute magnitude of A_f differs across normalizations because the reference category enters the subtraction, but the diagnostic ranking, which is what drives the substantive conclusions in Section 5.1, is preserved.

Table 12: Robustness indicators for the baseline swap from *Other* to *Sleep and Personal Care* as the reference activity. Larger values indicate stronger concordance for Top-K overlap, Jaccard, Spearman, and Kendall tau; smaller values indicate stronger concordance for max rank shift.

Indicator	Value
Top-5 overlap	5/5
Jaccard (Top-5)	1.00
Spearman rank correlation (all features)	0.879
Kendall tau (all features)	0.733
Max rank shift (Top-K union)	3

I.2 Covariate-Set Sensitivity

A natural concern is whether the diagnostic ranking is sensitive to which covariates enter the structural model. We test this by re-estimating the model under a leaner specification that removes one plausibly non-core covariate, weekly earnings, and keeps all other modeling choices identical. The lean specification therefore retains age, education, sex, marital status, and the four race indicators, plus an intercept. We re-estimate $\hat{\theta}_H$ and $\hat{\theta}_l$ for all five evaluated LLMs on the same lean design matrix, then recompute the feature-wise multi-model averaged deviation A_f on the intersection of features common to both specifications.

Table 14 reports the four robustness indicators. The Top-5 most misaligned features are identical across the two specifications, with a Jaccard index of 1.00. The Spearman rank correlation across the nine common features is 0.883, the Kendall tau is 0.722, and the maximum rank shift within the Top-K union is only 2 positions. Table 15 provides a Top-9 side-by-side comparison: removing earnings produces only mild local reordering, and the same set of features dominates the ranking under both specifications. We additionally verify sign stability on the Human estimate across all 27 (activity, feature) cells: 26 of 27 cells retain the same sign under the lean specification, and the single nominal flip (*Race: Asian* on *Sleep and Personal Care*) occurs at a coefficient that is statistically indistinguishable from zero in both specifications (Human estimate 0.003 in the full spec and −0.006 in the lean spec, with confidence intervals straddling zero in both cases). Taken together, the rankings, magnitudes,

Table 13: Top-10 features ranked by multi-model averaged deviation A_f, under the two equivalent normalizations. The same ten features appear in both rankings; differences in absolute A_f magnitudes are expected because the reference category enters the subtraction.

	Reference = Other (paper)		Reference = Sleep and PC	
Rank	Feature	A_f	Feature	A_f
1	Race: Pacific Islander	0.374	Race: Pacific Islander	0.273
2	Race: Native American	0.223	Sex	0.132
3	Race: Black	0.194	Spouse: Present	0.108
4	Spouse: Present	0.171	Race: Native American	0.099
5	Sex	0.098	Race: Black	0.092
6	Age	0.095	Age	0.074
7	Earnings	0.076	Race: Asian	0.062
8	Race: Asian	0.072	Earnings	0.052
9	Spouse: Unmarried	0.070	Spouse: Unmarried	0.040
10	Education	0.062	Education	0.033

and signs are stable, indicating that the diagnostic conclusions in Section 5.1 are not driven by the inclusion of any single covariate.

I.3 Corner Solutions and ε Smoothing

A subtle concern is that the utility specification in Eq. 4 involves $\ln(h_{ij})$, which is undefined at zero allocations. In our implementation, the softmax baseline parameterization avoids this issue at the model side, since predicted shares always lie in the open interval $(0, 1)$. However, observed shares can be exactly zero when an individual or LLM allocates no time to a given activity, and these corner solutions could in principle influence the recovered parameters through the residual term. We therefore re-estimate the model after a minimal ε smoothing of observed shares.

Smoothing protocol. For each respondent and each LLM output, any cell with zero share is replaced by $\varepsilon = 1/1440$, the share-level equivalent of one minute, and the row is then re-normalized to sum to one. Because one minute is 0.07% of the daily time budget, the perturbation is small by construction. We apply the same smoothing uniformly to the Human ATUS data and to all five LLM outputs so that the comparison $|\hat{\theta}_l - \hat{\theta}_H|$ is computed under identical preprocessing on both sides.

Zero-cell distribution. Table 16 reports where zero entries occur. The phenomenon is highly asymmetric: 44.7% of human respondents report zero *Work* minutes on the survey day, reflecting non-work days, weekends, and retirement, while LLM outputs almost never produce zero *Work* allocations (mean 0.4% across the five models). Zero entries in *Leisure* or *Sleep and Personal Care* are rare on both sides. This asymmetry is itself a mechanism-level signal: LLMs tend to produce "template" schedules that always allocate some positive time to each major activity, attenuating the long left tail observed in real human time use.

Robustness result. Table 17 reports the comparison between estimates with and without smoothing. The ranking of features by A_f is essentially unchanged: Spearman rank correlation is 1.000, Top-5 overlap is $5/5$, and the maximum rank shift on the Top-K union is 0. The full Top-10 side-by-side comparison in Table 18 shows that every feature retains its position and A_f values agree to three decimal places. We conclude that the diagnostic conclusions are not driven by corner solutions in the observed data.

I.4 Elasticity Sensitivity: τ Sweep

The utility specification in Eq. 4 corresponds to a Cobb-Douglas form with activity weights $\exp(\theta_j^\top X_i)$ and unit elasticity of substitution. To test whether the diagnostic ranking is sensitive to this assumption, we introduce a temperature parameter τ into the softmax share function,

$$s_{ij}(\tau) \;=\; \frac{\exp(\theta_j^\top X_i/\tau)}{\sum_k \exp(\theta_k^\top X_i/\tau)},$$

and sweep $\tau \in \{0.5, 0.8, 0.9, 1.0, 1.1, 1.2, 1.5\}$, where $\tau = 1.0$ recovers the paper specification. Smaller τ pushes allocations toward concentrated decisions, while larger τ flattens them. While τ is not formally identical to a CES elasticity parameter, it plays an analogous role in controlling how sharply the optimal allocation concentrates around its argmax, and therefore provides a practical CES-style perturbation of the underlying decision sharpness.

Table 14: Robustness indicators for the covariate-set sensitivity check that drops *Earnings* from the design matrix. Rank correlations are computed on the nine features common to both specifications.

Indicator	Value
Top-5 overlap	$5/5$
Jaccard (Top-5)	1.00
Spearman rank correlation (common features)	0.883
Kendall tau (common features)	0.722
Max rank shift (Top-K union)	2
Sign consistency on Human $\hat{\theta}$	$26/27 = 96.3\%$

Table 15: Top-9 features ranked by multi-model averaged deviation A_f under the full and lean specifications. *Earnings* appears only under the full specification because it is removed in the lean run; the remaining nine features form the common set used for the rank comparison.

	Full specification (paper)		Lean specification (no earnings)	
Rank	Feature	A_f	Feature	A_f
1	Race: Pacific Islander	0.374	Race: Pacific Islander	0.340
2	Race: Native American	0.223	Spouse: Present	0.212
3	Race: Black	0.194	Race: Black	0.201
4	Spouse: Present	0.171	Race: Native American	0.190
5	Sex	0.098	Sex	0.150
6	Age	0.095	Age	0.104
7	Earnings	0.076	Spouse: Unmarried	0.069
8	Race: Asian	0.072	Education	0.066
9	Spouse: Unmarried	0.070	Race: Asian	0.053

Estimation protocol. For each τ we re-estimate $\hat{\theta}_H$ and $\hat{\theta}_l$ for all five LLMs under the same τ, then recompute the feature-wise multi-model averaged deviation A_f. We do not constrain the scale of $\hat{\theta}$ across τ values: the optimizer is free to absorb the temperature into the magnitude of $\hat{\theta}$, so absolute coefficient sizes vary with τ while the relative ordering, which is what the rank-based diagnostic depends on, is preserved.

Robustness result. Table 19 reports the comparison between each τ and the base $\tau = 1.0$. The Top-5 most misaligned features form an identical set across all seven values of τ, with Jaccard index 1.00 throughout. Spearman rank correlations remain at 0.95 or above across the full sweep, and the maximum rank shift within the Top-K union is at most 2 positions. Table 20 provides the explicit Top-5 ranking at each τ: *Race: Pacific Islander* is always rank 1 and *Sex* is always rank 5, while the small reorderings at ranks 2 through 4 occur among three features (*Race: Native American*, *Race: Black*, *Spouse: Present*) whose A_f values are already close in the base specification, consistent with a near-tie boundary rather than a substantive instability. We conclude that the diagnostic ranking is robust to CES-style perturbations of the elasticity assumption.

I.5 Bootstrap Uncertainty for Feature-Wise Deviation

Feature-level deviation rankings can in principle be inflated by sampling noise from small-N subgroups. To quantify this, we run a non-parametric bootstrap over ATUS respondents with $B = 200$ replicates and seed 45. Each replicate resamples $N = 4{,}307$ respondents with replacement, re-estimates $\hat{\theta}_H^{(b)}$ on the resampled human data, and recomputes the feature-wise multi-model averaged deviation $A_f^{(b)}$. LLM estimates are held fixed throughout, so this design isolates sampling variability on the human side; LLM-side uncertainty is quantified separately below. We report 95% percentile intervals for each A_f and the probability that each feature appears in the Top-5 deviation set across replicates.

Power classification. The bootstrap precision of a race or partnership indicator depends primarily on the size of the subgroup it identifies, so we record the within-group count N_g for these indicators in addition to the full sample size N. We label a feature as *well-powered* when $N_g \geq 50$ and *exploratory* otherwise. Continuous and majority binary covariates such as age, education, earnings, and sex are estimated on the full sample and are well-powered by construction. This labeling makes the two small subgroup indicators (*Race: Pacific*

Table 16: Zero-cell distribution by decision maker and activity. Counts indicate the number of respondents with exactly zero allocation to the named activity; percentages are over total respondent-activity cells. Zero entries are concentrated in human *Work*, reflecting non-work survey days.

Decision maker	N	Total zero	%	Work	Leisure	Other
Human	4,307	2,220	12.9	1,927	284	7
GPT-4o	4,307	12	0.07	12	0	0
Claude-3.7	4,307	20	0.12	20	0	0
DeepSeek-V3	4,295	20	0.12	20	0	0
Llama-3.3	4,302	31	0.18	26	0	5
Qwen-2.5	4,301	12	0.07	4	0	8

Table 17: Robustness indicators for the ε-smoothing check. Smoothing leaves the diagnostic ranking essentially unchanged.

Indicator	Value
Top-5 overlap	5/5
Jaccard (Top-5)	1.000
Spearman rank correlation (all features)	1.000
Kendall tau (all features)	1.000
Max rank shift (Top-K union)	0

Islander with $N_g = 8$ and *Race: Native American* with $N_g = 34$) explicit as exploratory, consistent with the small-sample caveat already noted in Section 5.1 when omitting them from the attribute-level visualization.

Result. Table 21 reports the full bootstrap summary. The two exploratory features have very wide intervals: *Race: Pacific Islander* has 95% PI $[0.170, 1.154]$, a width on the order of unity, confirming that its apparent rank-1 magnitude is dominated by estimation noise rather than a stable signal. The two well-powered features that anchor the paper's substantive findings have tight intervals and near-deterministic Top-5 inclusion: *Spouse: Present* ($N_g = 2{,}333$) has PI $[0.145, 0.222]$ with Top-5 inclusion 1.00, and *Race: Black* ($N_g = 460$) has PI $[0.125, 0.288]$ with inclusion 0.99. We conclude that the headline conclusions in Section 5.1, which focus on *Race: Black* and *Spouse: Present*, are robust to sampling variability, while the exploratory race categories should be interpreted accordingly and not treated as ranked findings on their own.

LLM-side uncertainty. LLM coefficients carry their own estimation uncertainty, and decoding at temperature 0.1 is not strictly deterministic. We therefore run the mirror-image bootstrap ($B = 100$): we resample respondents, re-estimate $\hat{\theta}_l$ on the resampled LLM outputs, and hold the human estimate fixed. Across the five models, three activities, and the two headline features (*Race: Black* and *Spouse: Present*), the bootstrap standard deviations of the LLM coefficients range from 0.008 to 0.028, whereas the corresponding human–LLM gaps range from 0.04 to 0.31. The gap exceeds the LLM-side standard deviation by more than a factor of five in every headline cell except Llama-3.3 on *Race: Black* $\times$ *Work* (factor 1.4), which is also non-significant in the Wald tests of Appendix F; the two other cells that are non-significant in those tests are driven by human-side rather than LLM-side uncertainty. The scope of generation variability is further bounded by the strong concentration of LLM outputs documented in Appendix I.6: for example, GPT-4o produces only 15 distinct nonzero allocation values across all 17,228 generated entries.

I.6 Round-Number Heaping and Granularity

LLMs are known to favor round numbers when generating quantities, so part of the measured human–LLM difference could in principle reflect output granularity rather than trade-off weights. Table 22 quantifies this directly. Between 58.7% and 100% of nonzero LLM allocation entries are multiples of 30 minutes, versus 32.1% for humans, and each LLM uses only 15 to 45 distinct nonzero allocation values across all 4,307 profiles and four activities, versus 902 for humans. The statistics capture both the round-number preference and the concentrated "template" schedules visible in Figures 6 and 7. The latter is not a confound to be removed but part of the misalignment signal, since insensitivity of outputs to covariates is precisely what attenuated LLM coefficients measure.

Robustness result. To test whether granularity drives the diagnostics, we round every human allocation to the nearest 30 minutes, so that both sides share the same granularity, re-estimate $\hat{\theta}_H$,

Table 18: Top-10 features ranked by multi-model averaged deviation A_f, with and without ε smoothing. The ranking and the A_f values are identical up to the third decimal place.

	No smoothing (paper)		$\varepsilon = 1/1440$ smoothing	
Rank	Feature	A_f	Feature	A_f
1	Race: Pacific Islander	0.374	Race: Pacific Islander	0.373
2	Race: Native American	0.223	Race: Native American	0.223
3	Race: Black	0.194	Race: Black	0.195
4	Spouse: Present	0.171	Spouse: Present	0.171
5	Sex	0.098	Sex	0.098
6	Age	0.095	Age	0.095
7	Earnings	0.076	Earnings	0.076
8	Race: Asian	0.072	Race: Asian	0.072
9	Spouse: Unmarried	0.070	Spouse: Unmarried	0.070
10	Education	0.062	Education	0.062

Table 19: Robustness indicators across the τ sweep. Each row compares the ranking at the given τ to the base specification at $\tau = 1.0$. The Top-5 set is identical at every τ, and rank correlations remain at 0.95 or above throughout.

τ	Top-5 overlap	Jaccard	Spearman	Kendall τ	Max shift
0.5	5/5	1.00	0.952	0.867	1
0.8	5/5	1.00	0.964	0.911	0
0.9	5/5	1.00	0.952	0.867	2
1.0	5/5	1.00	1.000	1.000	0
1.1	5/5	1.00	0.988	0.956	0
1.2	5/5	1.00	0.988	0.956	0
1.5	5/5	1.00	0.988	0.956	0

and recompute A_f. Table 23 shows that the feature ranking is unchanged at every position, and the human coefficients on *Race: Black* and *Spouse: Present* move by at most 0.003. Together with the ε-smoothing check in Appendix I.3, which addresses the corner-solution side of the same concern, this indicates that the diagnostic conclusions are not artifacts of numerical generation patterns.

I.7 Diary Day versus Typical Day

ATUS records activities for one designated diary day, whereas the LLM prompt asks for a typical day. Merging the IPUMS diary-day variable shows that this difference matters for the human benchmark, because ATUS oversamples weekend diaries by design: 48.4% of our analytic sample are weekend days (versus $2/7 \approx 28.6\%$ of a calendar week), and the zero-work rate is 73.1% on weekend diaries versus 18.1% on weekday diaries. Much of the zero-work asymmetry in Table 16 therefore reflects sample composition.

Estimand-matched reweighting. A typical day corresponds to a day-of-week-unconditional expectation, i.e., a mixture of $5/7$ weekday and $2/7$ weekend days. We re-estimate $\hat{\theta}_H$ with observation weights that reproduce this mixture, holding the LLM estimates fixed. Table 24 shows that the diagnostic conclusions are preserved: the Top-5 set is identical, the Spearman rank correlation is 0.939, and the only rank change within the Top-K union is a swap between the two exploratory race categories. The headline coefficients keep their signs: *Race: Black* remains positive and *Spouse: Present* negative in all three activities.

Weekday decomposition. Re-estimating on weekday diaries only ($N = 2{,}221$) reveals genuine day structure in human behavior: the *Race: Black* coefficients on *Sleep and Personal Care* and *Work* roughly halve, indicating that these human differences concentrate on weekends, while *Spouse: Present* remains negative across all activities. The ranking correspondingly changes more (Table 24), partly because the subsample is half the size. We report this decomposition because it exposes an additional dimension of misalignment: LLM outputs elicited without day information are day-invariant by construction and cannot reproduce this weekday/weekend structure. Since a typical day is itself ambiguous between an average day and a representative weekday, the two re-estimations correspond to these two readings.

J Out-of-Domain Validation on HH-RLHF

The main paper instantiates XCHOICE on a continuous-share decision (daily time allocation). To probe whether the framework extends beyond this setting, we apply XCHOICE to a standard NLP discrete-preference dataset, Anthropic HH-RLHF, in which each datum is a binary pairwise preference between a chosen response y^+ and a rejected response y^-.

Table 20: Top-5 features ranked by multi-model averaged deviation A_f at each τ. The set of Top-5 features is identical across all τ, and the only reorderings occur at ranks 2 through 4 among three features whose A_f values are close in the base specification.

Rank	τ=0.5	τ=0.8	τ=0.9	τ=1.0	τ=1.1	τ=1.2	τ=1.5
1	Race: Pac. Isl.	Race: Pac. Isl.	Race: Pac. Isl.	Race: Pac. Isl.	Race: Pac. Isl.	Race: Pac. Isl.	Race: Pac. Isl.
2	Race: Nat. Am.	Race: Nat. Am.	Race: Black	Race: Nat. Am.	Race: Nat. Am.	Race: Nat. Am.	Race: Nat. Am.
3	Spouse: Present	Race: Black	Spouse: Present	Race: Black	Race: Black	Race: Black	Race: Black
4	Race: Black	Spouse: Present	Race: Nat. Am.	Spouse: Present	Spouse: Present	Spouse: Present	Spouse: Present
5	Sex	Sex	Sex	Sex	Sex	Sex	Sex

Table 21: Bootstrap uncertainty for feature-wise deviation ($B = 200$, seed 45, $N = 4{,}307$ respondents per replicate). The point estimate is the multi-model averaged $|\hat{\theta}_l - \hat{\theta}_H|$ on the original sample. Bootstrap median and 95% percentile intervals are taken over replicates. Top-5 inclusion is the fraction of replicates in which the feature appears in the top five by A_f. N_g is the within-group count for race and partnership indicators (the relevant precision-determining sample) and "–" for continuous and majority binary covariates that are estimated on the full sample. Features with $N_g < 50$ are labeled exploratory.

Feature	Point A_f	Median	PI 2.5%	PI 97.5%	Top-5 incl.	N_g	Class
Race: Pacific Islander	0.374	0.423	0.170	1.154	1.000	8	exploratory
Race: Native American	0.223	0.257	0.197	0.438	1.000	34	exploratory
Race: Black	0.194	0.205	0.125	0.288	0.990	460	well-powered
Spouse: Present	0.171	0.177	0.145	0.222	1.000	2,333	well-powered
Sex	0.098	0.097	0.086	0.107	0.175	–	well-powered
Age	0.095	0.098	0.086	0.112	0.260	–	well-powered
Earnings	0.076	0.075	0.060	0.095	0.005	–	well-powered
Race: Asian	0.072	0.088	0.070	0.142	0.210	294	well-powered
Spouse: Unmarried	0.070	0.090	0.040	0.194	0.360	249	well-powered
Education	0.062	0.061	0.045	0.085	0.000	–	well-powered

Table 22: Round-number heaping by decision maker, computed over nonzero allocation entries (four activities per profile).

Decision maker	% mult. of 30	% mult. of 60	Distinct values
Human	32.1	18.5	902
GPT-4o	100.0	100.0	15
Claude-3.7	98.6	87.0	33
DeepSeek-V3	100.0	98.5	28
Llama-3.3	99.2	99.0	45
Qwen-2.5	58.7	58.6	40

Table 23: Robustness indicators for the granularity check, in which human allocations are rounded to the nearest 30 minutes before re-estimation.

Indicator	Value
Top-5 overlap	5/5
Jaccard (Top-5)	1.00
Spearman rank correlation (all features)	1.000
Kendall tau (all features)	1.000
Max rank shift (Top-K union)	0
Max change in headline human coefficients	0.003

Instantiation. We map each response to a vector of interpretable text features $\phi(y) \in \mathbb{R}^K$, form the contrast $\Delta\phi = \phi(y^+) - \phi(y^-)$, and recover the preference vector θ via the logistic choice model

$$P(y^+ \succ y^- \mid x) \;=\; \sigma(\theta^\top \Delta\phi),$$

which is the natural discrete analogue of the optimal decision rule in Section 3. We use $K = 13$ lightweight, fully interpretable features grouped into five categories: length (*len_chars*, *len_tokens*), structure and formatting (*num_sent_punct*, *num_newlines*, *num_bullets*, *has_numbered_list*, *has_colon_newline*), punctuation (*num_questions*, *num_exclaims*), lexical politeness and stance markers (*has_polite_marker*, *has_hedging_marker*, *has_refusal_marker*), and a coarse safety signal (*safety_word_count*). Each ϕ entry is standardized before fitting; estimation uses regularized maximum likelihood with ℓ_2 penalty on 80,000 training pairs from the HH-RLHF train split, evaluated on the full test split of 8,552 pairs. We add a symmetric copy of every pair with the sign of $\Delta\phi$ flipped, so that the classifier sees both directions of the comparison and the intercept is identically zero. Bootstrap percentile intervals are computed over $B = 100$ resamples of the training pairs.

Recovered preference structure. Table 25 reports the recovered θ vector, sorted by absolute magnitude. Twelve of the thirteen features have 95% percentile intervals that exclude zero, indi-

Table 24: Diagnostic ranking under diary-day adjustments of the human benchmark, compared with the unweighted full-sample estimates.

Human sample	Top-5	Jaccard	Spearman	Kendall τ	Max shift
Typical-day reweighting	$5/5$	1.00	0.939	0.822	1
Weekday only	$3/5$	0.43	0.685	0.556	5

Table 25: Recovered preference parameters θ on HH-RLHF, sorted by absolute magnitude. Features are standardized before fitting. Bootstrap 95% percentile intervals are computed over $B = 100$ resamples on 80,000 training pairs. "Sig." indicates whether the 95% PI excludes zero.

Feature	θ	PI 2.5%	PI 97.5%	Sig.
len_chars	0.623	0.493	0.715	✓
len_tokens	-0.594	-0.682	-0.456	✓
has_polite_marker	0.084	0.074	0.094	✓
has_colon_newline	-0.059	-0.071	-0.048	✓
has_hedging_marker	-0.056	-0.066	-0.045	✓
has_refusal_marker	0.051	0.041	0.061	✓
num_newlines	0.041	0.026	0.055	✓
num_questions	-0.040	-0.051	-0.029	✓
safety_word_count	-0.040	-0.050	-0.027	✓
num_exclaims	-0.027	-0.038	-0.018	✓
has_numbered_list	0.020	0.010	0.029	✓
num_bullets	0.011	0.001	0.023	✓
num_sent_punct	0.007	-0.012	0.028	–

cating clear directional preferences. Length features dominate: *len_chars* has $\theta = 0.623$ (positive) and *len_tokens* has $\theta = -0.594$ (negative). Their opposite signs reflect the fact that, after standardization, the two are highly collinear surrogates for response length; the logistic optimum splits the joint signal between them, and what matters substantively is that long responses are preferred on net. Beyond length, several smaller but statistically significant signals match patterns documented in the RLHF literature: a positive coefficient on *has_polite_marker* ($\theta = 0.084$) and on *has_refusal_marker* ($\theta = 0.051$, consistent with HH-RLHF training for harmlessness in addition to helpfulness), and negative coefficients on *has_hedging_marker* ($\theta = -0.056$) and *safety_word_count* ($\theta = -0.040$).

Outcome metric vs. mechanism. The fitted logistic model attains 52.9% accuracy and ROC-AUC 0.534 on the test set. By itself this is only modestly above chance, and a purely outcome-based assessment would suggest the framework provides little signal. However, the recovered θ vector tells a different story: it is statistically significant on twelve of thirteen interpretable dimensions and recovers familiar preference patterns from the RLHF literature. This contrast underscores the central argument of the paper, stated in Section 1: outcome metrics such as accuracy can be uninformative for diagnosing decision mechanisms. XCHOICE provides interpretable, directionally consistent preference parameters in a setting where outcome accuracy alone would be uninformative, and therefore extends naturally beyond the ATUS continuous-share instantiation to standard discrete-preference benchmarks.