PEMAND: Persona-Enriched Multi-Agent Negotiation for Household Decision-Making

Anonymous ACL submission

Abstract

Modeling household-level decisions is central to many real-world applications, including trip planning, residential mobility and migration, disaster management, etc. Existing studies primarily rely on classical machine learning models with limited predictive capacity, while recent LLM-based approaches have yet to incorporate behavioral theory or intra-household interaction dynamics, both of which are essential for modeling realistic household decisions. To address these limitations, we propose Persona-Enriched Multi-Agent Negotiation for household Decision-making (PEMAND), a novel LLM-based framework that integrates behavioral theory into individualized, household-aware persona modeling and simulates household-level decision-making through structured multi-agent negotiation. Specifically, PEMAND transforms static sociodemographic attributes into coherent narrative profiles that explicitly encode household-level attitudes, subjective norms, and perceived behavioral controls, following our proposed Household-Aware Chain-of-Planned-Behavior (HA-CoPB) framework. Building on these theory-grounded personas, PEMAND captures real-world household decision negotiation via a structured two-phase multi-agent conversation framework with a novel persona-alignment control mechanism. Evaluated on national and regional household decision datasets across two major domains, including travel behavior and residential mobility, PEMAND consistently outperforms state-of-the-art benchmarks.

1 Introduction

Household decision-making is fundamental in real-world applications such as transportation planning (Talpur et al., 2023), urban development (Wang et al., 2021), emergency management (Cova et al., 2024), and retail (Erasmus et al., 2014). However, modeling these decisions remains challenging for AI and machine learning (AI/ML), as household decision-making emerges from dynamic interactions among members' preferences, roles, constraints, and multi-stage cognitive processes (Ajzen, 1991; Wang and Ruhe, 2007; Golob and McNally, 1997; Yang et al., 2016; Leyer and Freytag-Leyer, 2025). This creates a critical need for novel methodological frameworks that can capture such complexities, to improve predictive accuracy and support policy and planning across diverse decision domains (Acosta et al., 2020).

Prior machine learning approaches for modeling household decisions (Sun et al., 2024a; Zhang et al., 2019; Xue and Yao, 2022; Firoozzare et al., 2024; Amarasinghe Arachchige et al., 2022) often treat the household as a monolithic unit or a simple aggregation of individuals, failing to capture complex intra-household interactions (Dong and Kaiser, 2008; Anand et al., 2026; Mu et al., 2026).

While large language models (LLMs) can leverage unstructured information and produce flexible representations of household contexts and decision factors (Xiao and Wang, 2025; Chen et al., 2025; Lim et al., 2026), and multi-agent systems make it possible to simulate intra-household interaction, coordination, and negotiation dynamics (Abdelnabi et al., 2023, 2024), they face three critical limitations when applied to decision prediction. *First*, existing LLM frameworks lack explicit grounding in multi-stage cognitive theories (Ajzen, 1991; Wang and Ruhe, 2007), and even when enriched with sociodemographic and contextual information, they may partially capture the underlying decision-making mechanisms without behavioral guidance (Chen et al., 2025; Tjuatja et al., 2024). *Second*, they rely on generic dialogue protocols rather than structured communication that reflects real-world household deliberation processes, where members exchange preferences, coordinate constraints, and resolve conflicts before reaching a collective outcome (Golob and McNally, 1997; Kong et al., 2025). *Third*, although LLMs provide expressive

representations and strong reasoning capabilities, they remain prone to hallucinations (Bang et al., 2025), invalid or inconsistent outputs (Ahn and Yin, 2025), and misalignment with observed human behavior (Lei et al., 2024; Wang et al., 2023).

To address these limitations, we propose **Persona-Enriched Multi-Agent Negotiation for household Decision-Making (PEMAND)**, a novel two-stage LLM-based framework that bridges theory-grounded cognitive modeling with collective household negotiation. Our approach simulates household decisions by assigning each LLM agent the role of a household member (e.g., mother, father, child) and guiding the decision process through two phases. First, we synthesize factually-grounded personas using a heuristic anchoring mechanism (Tversky and Kahneman, 1974) to ensure agents reflect specific sociodemographic distributions rather than generic stereotypes. Specifically, we extend the Chain-of-Planned-Behavior (CoPB) framework (Shao et al., 2024) into a Household-Aware CoPB (HA-CoPB) formulation grounded in the Theory of Planned Behavior (TPB) (Ajzen, 1991), a well-established framework for modeling human decision-making. HA-CoPB adapts TPB components to the household setting by representing Perceived Behavioral Control (PBC) as shared resource constraints, such as vehicle availability, and Subjective Norms (SN) as concrete intra-household dependencies. To improve persona reliability, we further introduce an Alignment via Refinement (AvR) pattern (Madaan et al., 2023), where a review agent critiques the generated reasoning against historical priors (Tversky and Kahneman, 1974) and household constraints, prompting the primary agent to iteratively revise its behavioral profile. In the second stage, these agents engage in a structured negotiation protocol to resolve conflicting needs and converge on a realistic, consensus-based household decision. This process consists of two phases: (1) a parallel proposal phase that captures heterogeneity in individual intentions, and (2) a consensus refinement phase that models intra-household interaction and collective decision-making. We further introduce an alignment control mechanism consisting of a human-calibrated LLM-judge moderator, persona supervised fine-tuning (SFT), and negotiation reinforcement self-training (ReST). This mechanism encourages each agent to adhere to its assigned HA-CoPB, interact validly with other household agents, and produce decisions that better align with observed real-world behavior.

Through evaluations from two distinct domains, including trip planning and relocation decisions, using national and regional scale survey datasets, we compare PEMAND with a suite of state-of-the-art baselines and show that PEMAND consistently achieves superior performance across datasets.

Our primary contributions include:

- **A Novel Multi-stage Methodological Framework for Household Decision Modeling.** We introduce a new multi-stage formulation for household decision prediction that treats household decisions as collectively negotiated outcomes rather than independent individual predictions. Unlike prior LLM-based decision models based on static personas or direct prompting, PEMAND unifies household-aware persona construction, role-specific proposal generation, constraint-aware negotiation, and alignment control in a single inference pipeline.
- **Behavioral Theory-Guided Persona Inference.** We introduce a theory-grounded multi-agent architecture that embeds TPB-based cognitive reasoning into household simulation. By adapting TPB to shared household constraints and enforcing a hierarchical priority mechanism, PEMAND anchors each agent's intentions in constrained HA-CoPB profiles before negotiation, reducing unconstrained hallucination beyond static persona-based prompting.
- **Structured Multi-agent Negotiation with Alignment Control.** We introduce a structured household negotiation and alignment mechanism that models collective decisions as a controlled deliberation process rather than generic multi-agent dialogue. The alignment module uses human calibration to ground the LLM judge in expert-validated behavioral criteria, providing an external alignment signal beyond model self-assessment.

2 Related Work

Modeling Household Decision. Household decision-making has been widely studied across multiple domains, including transportation planning (Anand et al., 2026, 2025), residential mobility (Wang et al., 2021; Clark and Lisowski, 2017), emergency management (Sun et al., 2024a; Cova et al., 2024), and retail marketing (Erasmus et al., 2014; İnan and Konyalı, 2025). Existing studies have employed a range of modeling approaches to analyze and predict household decisions, in-

cluding statistical models (Clark and Lisowski, 2017; Al Kurdi et al., 2025; Kuligowski et al., 2022; Qawasmeh, 2024), machine learning models (Anand et al., 2026; Sun et al., 2024a; Naseralavi et al., 2025; Qawasmeh, 2024), economic models (Buchmann et al., 2025), and recently, LLM models (Chetty et al., 2025; Pijp). Existing models often treat households as single decision-making units or aggregate individual choices, overlooking intra-household interaction dynamics (Mu et al., 2026; Dong and Kaiser, 2008; Anand et al., 2026). Even theory-grounded models (Kuligowski et al., 2022; Sun et al., 2024b) still struggle to capture realistic multi-stage decision processes.

Persona Generation for Decision Modeling. Persona generation for human decision modeling has shifted from static demographic profiles to richer, psychologically grounded LLM-based representations. Recent work uses LLMs to infer latent cognitive antecedents (Shao et al., 2024; Sameen et al., 2025), synthesize daily activity schedules from aggregate data (Amin et al., 2025), and forecast long-term life-course decisions such as residential mobility from textual trajectories and panel datasets (Stark et al., 2025; Gao et al., 2026).

Despite this progress, key gaps remain for joint household-level prediction. Standard zero-shot initialization can suffer from population misalignment and persona degradation during multi-agent dialogue (Hu et al., 2025; Frisch and Giulianelli, 2024). Existing isolated generation methods also miss intra-household entanglement, including shared resources, relational constraints, and interpersonal obligations (Liu et al., 2025). Moreover, independent personas often lack the communicative agency and constraint enforcement needed for structured negotiation, leading to implausible individual outcomes and weak collective consensus.

LLM-based Multi-agent Conversation. Recent work shows that communication among LLM-based agents can improve collective task performance (Zhang et al., 2024; Yan et al., 2025; Wu et al., 2024). A common paradigm assigns agents explicit roles or personas, enabling structured collaboration and more realistic role-playing interactions (Abbasiantaeb et al., 2024; Wu et al., 2024; Li et al., 2023; Feng et al., 2025). Richer persona specifications, including background, emotional state, and psychological traits, further improve behavioral alignment in dialogue (Shao et al., 2023; Yang et al., 2025; Wu et al., 2025). However, most role-playing multi-agent frameworks still rely on prompt-based persona conditioning without explicit mechanisms for sustained alignment with real-world interaction dynamics, limiting their use in behaviorally grounded decision-making.

3 Methodology

We propose a two-stage LLM-based framework for household decision prediction. Rather than directly mapping household attributes to outcomes, our framework models the intermediate processes through which household decisions are formed, including individual reasoning and intra-household negotiation. Given a household $\mathcal{H}$ with N members $\{m_1, m_2, \ldots, m_N\}$, along with their demographic and contextual attributes, the goal is to predict the final household-level decision, such as trip count or relocation choice. We formulate this task as a multi-stage process in which each member first generates a persona-grounded individual proposal, after which members engage in structured negotiation to reach a collective decision. Figure 1 provides an overview of the proposed framework.

3.1 Behavioral Theory-Guided Persona Generation

Our persona generation module bridges static demographic data and dynamic behavioral reasoning by integrating deterministic fact extraction with constrained generative reasoning. This approach ensures that the synthesized narrative is factually accurate and structurally consistent with the overall household composition. More importantly, the framework is domain-agnostic and can be instantiated for different household-level decision tasks.

3.1.1 Narrative Synthesis and Behavioral Anchor

Deterministic Fact Extraction. We first employ a deterministic *Translation Map* ($\Phi : \mathcal{X} \rightarrow \mathcal{F}$) that converts encoded survey responses into atomic natural language facts (e.g., $\Phi(\texttt{EDUC} = 5)$ maps to "*I have a graduate or professional degree.*"). This step utilizes a non-generative approach to ensure data fidelity. We also apply a *Neutralization Filter* at this stage to map ambiguous non-response codes to generic descriptors, preventing downstream hallucination of unfounded relationship dynamics.

Narrative Synthesis with Faithfulness Constraints. We then feed the atomic facts $\mathcal{F}$ into an instruction-tuned LLM to synthesize a first-person narrative $\mathcal{P}$. To ensure data faithfulness and mitigate sociodemographic hallucination (Ji

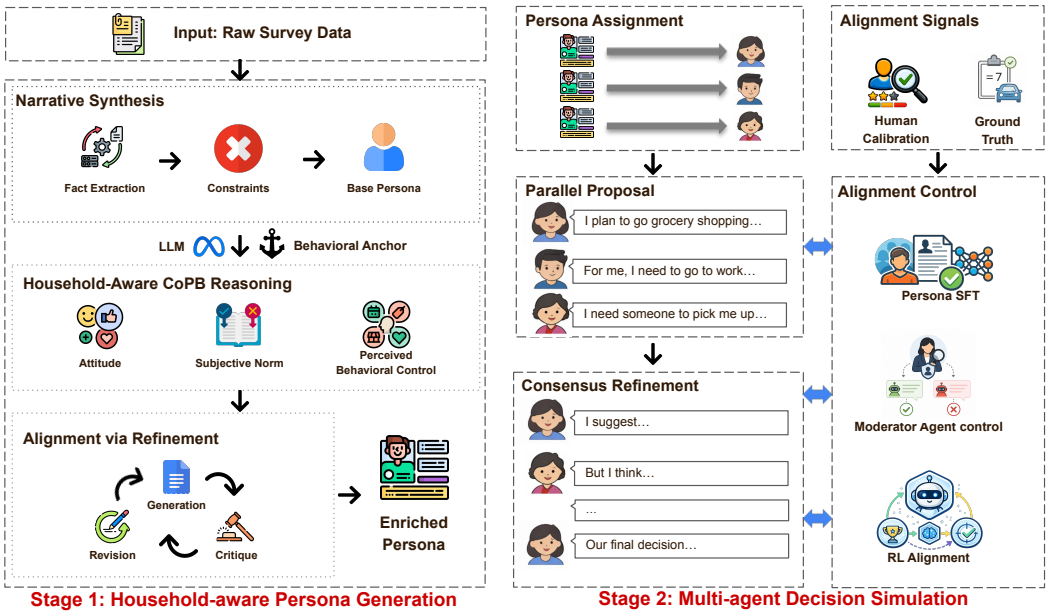

Figure 1: Methodological Framework for PEMAND.

et al., 2023), the system prompt imposes strict *Consistency Constraints* and task-specific *Vocabulary Restrictions*. This guarantees a clinically neutral but cognitively coherent persona profile that serves as the agent's implanted memory, aligned with its ground-truth attributes (detailed in Appendix D).

Behavioral Anchor (Historical Prior). To ground the LLM's reasoning in statistical reality and prevent data leakage (Hyndman and Athanasopoulos, 2018), we employ a historical **Behavioral Anchor** mechanism. We guide the LLM using a population-level reference point derived strictly from past survey cycles. This operationalizes the *Anchoring and Adjustment Heuristic* (Tversky and Kahneman, 1974) within a Bayesian framework (Washington et al., 2020), where the model treats the historical anchor as a prior and updates it based on specific persona constraints.

3.1.2 Household-Aware CoPB (Reasoning Engine)

Our *Household-Aware CoPB (HA-CoPB)* framework models the agent's cognitive process as a generative transformation, $f_{\text{HA-CoPB}}$, mapping the static persona to an **Enriched Persona** ($\mathcal{P}'$):

$$\mathcal{P}' = f_{\text{HA-CoPB}}(\mathcal{P}, \mathcal{H}_{ctx} \mid y_{\text{prior}}^{t-\text{lag}}) \quad (1)$$

where $\mathcal{H}_{ctx}$ represents the shared household context (e.g., resource bottlenecks, household member attributes). This formulation is grounded in the **Theory of Planned Behavior (TPB)** (Ajzen, 1991), a well-established theory of decision-making, and adapts its core psychological constructs to the household level (Appendix E.1):

$$\mathcal{P}' = \mathcal{P} \cup \{\mathcal{A}, \mathcal{SN}, \mathcal{PBC}\} \quad (2)$$

These constructs capture individual intrinsic utility ($\mathcal{A}$), intra-household obligations ($\mathcal{SN}$), and shared resource barriers ($\mathcal{PBC}$). To prevent hallucination, generation follows a strict Hierarchical Priority Mechanism ($\mathcal{PBC} \succ \mathcal{SN} \succ \mathcal{A}$), mandating that physical resource limits override social obligations and individual desires.

3.1.3 Alignment via Refinement (AvR) Mechanism

To synthesize the individual reasoning generated by the HA-CoPB framework into a robust psychological profile, we implement an **Alignment via Refinement (AvR)** architecture. Across all domains, the persona is finalized through an iterative dual-agent cycle. A primary reasoning agent (the Actor) leverages the HA-CoPB template to generate an initial persona narrative based on demographics and behavioral anchors. Subsequently, a secondary review agent (the Reviewer) evaluates this rationale, contrasting it against historical population anchors and structural household constraints (e.g., resource deficiencies). The Reviewer provides explicit critique, prompting the Actor to revise and calibrate its reasoning. This generation-critique-revision loop ensures that the resulting **Enriched Persona** is both theoretically grounded and behaviorally realistic, preventing unconstrained hallu-

cination before any downstream decision-making occurs.

3.2 LLM-based Multi-agent Conversation

3.2.1 Multi-agent Component

We design a simulation framework that captures realistic intra-household discussions for household decision-making, while incorporating additional control mechanisms to improve LLM alignment. The system comprises N+1 agents, A_1, A_2, $\cdots$, A_{N+1}, where N corresponds to the number of household members. The first N agents represent household members with distinct roles, sociodemographic attributes, and personas, while A_{N+1} serves as a moderator that enforces discussion-level constraints and supports coherent, aligned interaction dynamics.

Household Member Agent. Household member agents simulate participants in household discussion and collective decision-making. Each agent represents a distinct household role with corresponding behavioral patterns, preferences, and constraints. To link generated personas to negotiation behavior, every member agent is assigned HACoPB persona-specific attributes that guide its responses and interactions, as described below:

$$P_i = \{role_i, \mathcal{P}_i^{'}, \mathcal{H}_{ctx_i}\} \tag{3}$$

where $role_i$ specifies the agent's role within the household (e.g., parent or child) and indicates whether the agent serves as the household lead.

Moderator. The moderator is a non-member, human-calibrated LLM judge responsible for intra-conversation control (Appendix D.5). To ensure its evaluations reflect real-world decision standards, it is rigorously calibrated against human annotators (calibration metrics are detailed in Appendix H). It monitors the discussion using member attributes, shared household resources, and dialogue context. After each utterance, it checks whether the response is persona-consistent, constraint-compliant, discussion-relevant, and supported by a justified decision or proposal.

3.2.2 Multi-phase Multi-agent Household Discussion Framework

Prior work views household decision-making as shaped by both individual behaviors and household-level characteristics. Members form initial intentions from heterogeneous preferences and behavioral tendencies (Wang et al., 2024; Li et al., 2024), while final household travel decisions emerge through negotiation, compromise, and adaptation to shared resource constraints (Golob and McNally, 1997; Naseralavi et al., 2025).

We propose a two-phase interaction protocol for realistic household-level decision-making that captures both individual intention formation and intra-household negotiation. In the **parallel proposal phase**, members independently express their initial intentions; in the **consensus refinement phase**, agents iteratively deliberate and revise their proposals to reach a collective household decision.

Parallel Proposal. In the first phase, each household member agent independently generates an initial decision proposal conditioned on its persona profile (Appendix D.3). For agent A_i, the initial vote $v_i^{(0)}$ is sampled as:

$$v_i^{(0)} \sim \pi_{\theta_i}(\cdot \mid P_i). \tag{4}$$

Parallel generation elicits heterogeneous intentions while avoiding turn-order and anchoring biases. The individual votes are then aggregated into an initial household-level decision $\hat{y}^{(0)}$:

$$\hat{y}^{(0)} = G(v_1^{(0)}, v_2^{(0)}, \cdots, v_N^{(0)}), \tag{5}$$

where $G(\cdot)$ denotes a task-specific aggregation operator, such as summation for numerical decisions or majority voting for binary decisions. This aggregate serves as the starting point for the subsequent household conversation.

Consensus Refinement. In the consensus refinement phase, member agents deliberate over the household decision using a round-robin protocol for up to $R_{\max}$ rounds. At round t, each agent generates an utterance $u_i^{(t)}$ conditioned on its persona attributes P_i, the current household estimate $\hat{y}^{(t-1)}$, compact summaries of other members C_{-i}, and the conversation history h_{t-1} (Appendix D.4):

$$u_i^{(t)} \sim \pi_{\theta_i}(\cdot | P_i, \hat{y_i}^{(t-1)}, C_{-m}, h_{t-1}) \tag{6}$$

Each utterance may express agreement or disagreement with the current household-level estimate, introduce evidence-grounded revisions to the proposed decision, or negotiate compromises among members' preferences and constraints. After each round, we extract each speaker's preferred decision and apply revision clamps to numerical outputs to reduce sampling-induced fluctuations, while binary decisions remain unchanged. Details of the clamping procedure are provided in Appendix F. The

discussion ends when all speaker agents agree on a preferred decision or when $R_{\max}$ is reached. If no consensus is reached, the final household decision is computed from the speakers' final preferences using the median for numerical outputs or majority vote for binary outputs.

3.2.3 Role-aligned Control Mechanism

Reliable LLM alignment remains a key challenge in multi-agent household decision modeling. Prior work uses persona specifications to induce human-like, heterogeneous behavior (Amin et al., 2025; Wang et al., 2024), but persona conditioning alone cannot ensure profile consistency or realistic, coherent negotiation across multi-turn interactions (Wang et al., 2024).

To address these limitations, we introduce two role-aligned control mechanisms: intra-conversation control, where a moderator validates agent behavior during simulation, and out-of-conversation supervision, where role-conditioned SFT and ReST-based trajectory optimization ground member agents before deployment. To ensure reliable moderation, we also design task-specific rubrics and calibrate the LLM judge against human annotations.

Human Calibration. All LLM judges in the alignment control module are calibrated against human annotations to ensure that they evaluate responses according to expert-validated rubrics rather than model self-assessment. This calibration reduces potential judge-induced bias and improves the reliability of response validation. Calibration details are provided in Appendix H.

Intra-conversation Control. Under intra-conversation control, the moderator acts as a human-calibrated LLM judge that enforces response validity during generation. At each round t, it evaluates each utterance $u_i^{(t)}$ along four criteria: persona consistency, constraint adherence, interaction validity, and decision quality, with definitions in Appendix H. Utterances that fail to receive full scores on all criteria are discarded and regenerated before entering the conversation history, preventing misaligned or infeasible responses from influencing later interactions.

Out-of-conversation Supervision. In the parallel proposal phase, we fine-tune the LLM to generate persona-consistent and constraint-compliant initial decisions. We construct supervision from the training split using an iterative generation-and-filtering procedure: candidate proposals are sampled from the base model and evaluated by the human-calibrated LLM judge for persona consistency and constraint adherence. Only proposals receiving full scores on both criteria are retained as SFT targets, forming a role-specific proposal dataset $\mathcal{D}i^{prop} = (P_i, q\text{prop}, v_i^{(0)\star})$, where q_{prop} denotes the proposal prompt.

In the consensus refinement phase, we promote realistic and persona-consistent household negotiation using an inference-time best-of-N ReST trajectory selection strategy. For each training household, we sample N independent discussion trajectories and assign each a weighted reward score:

$$r = -\omega_{GT} \cdot |\hat{y} - y| + \omega_R \cdot \bar{\phi} \tag{7}$$

where $\hat{y}$ is the predicted household decision, y is the ground-truth decision, and $\bar{\phi}$ is the average score across the four PEMAND rubric dimensions from the human-calibrated LLM judge. For each household, we retain the highest-reward trajectory as high-quality self-generated supervision for SFT. At test time, the ReST-tuned policy is deployed under standard inference without access to ground-truth labels.

4 Experiment

4.1 Experiment Setting

Implementation Details We evaluate PEMAND under homogeneous backbone configurations, using **Llama-3.1-8B-Instruct** (Grattafiori et al., 2024) end-to-end for persona synthesis, household member agents, and the LLM judge in the main experiments. The judge is calibrated with human annotations and achieves strong reliability: Fleiss' κ (Fleiss, 1971) $= 0.620$ and QWK (Cohen, 1968) $= 0.698$ for travel, and $\kappa = 0.853$ and QWK $= 0.851$ for mobility. Full annotation details are provided in Appendix H. We prioritize offline open-weight models to support broad adoption, enable SFT, and protect sensitive demographic data. To assess robustness, we also evaluate **Gemma-3-27b-it** (Google DeepMind, 2025) and **GPT-5.4-mini** (OpenAI, 2026). Detailed parameter settings, computational costs, latency, and QLoRA fine-tuning hyperparameters are provided in Appendix A.2; robustness analyses are reported in Appendices C.1.

4.2 Main Results

Baselines. We benchmark PEMAND against conventional ML models, including Gradient Boosting and Random Forest, using identical features and

	Travel Domain (Regression)								Mobility Domain (Classification)			
	NHTS 2017				Puget Sound 2023				PSID 2021–2023			
Model	MAE↓	RMSE↓	sMAPE↓	±2 Acc↑	MAE↓	RMSE↓	sMAPE↓	±2 Acc↑	Acc↑	F1↑	Prec↑	Rec↑
Traditional ML												
Linear / Logistic Reg.†	3.34	4.52	51.00	0.40	2.83	3.86	58.54	0.58	0.69	0.55	0.47	0.67
Random Forest	3.16	4.36	51.74	0.43	2.75	3.74	57.17	0.59	0.72	0.51	0.51	0.51
Gradient Boosting	3.07	4.27	51.87	0.44	2.76	3.78	57.47	0.58	0.69	0.55	0.48	0.67
MLP	3.35	4.53	50.84	0.40	2.79	3.77	57.82	0.58	0.74	0.43	0.58	0.35
LLM Baselines												
Demographics-only	4.06	5.59	65.29	0.44	3.31	5.10	61.75	0.54	0.59	0.40	0.35	0.49
Persona	3.75	5.15	63.52	0.45	2.89	3.99	59.52	0.56	0.68	0.44	0.44	0.43
RAG	4.32	6.18	75.25	0.41	5.00	6.60	76.49	0.35	0.61	0.48	0.39	0.63
Self-consistency	3.74	5.11	63.66	0.46	2.89	4.04	58.30	0.57	0.43	0.45	0.31	**0.81**
PEMAND	**2.38**	**3.93**	**34.48**	**0.60**	**1.99**	**3.25**	**47.92**	**0.78**	**0.79**	**0.73**	**0.69**	0.78

Table 1: **Main Household Prediction Performance.** Comparison of ML and LLM baselines across travel-domain regression and mobility-domain classification. **Bold** denotes the best result within each dataset block. ↓/↑ indicate lower/higher is better. †Linear regression for NHTS/Puget Sound and logistic regression for PSID.

data splits. We also compare against four LLM baseline families: **Demographics-Only**, which predicts directly from socioeconomic features; **Persona**, which applies the HA-CoPB prompt to the full household in one inference step; **RAG**, which augments Persona with retrieved similar households; and **Self-Consistency**, which aggregates over $K = 10$ sampled predictions. Detailed prompts are provided in Appendix D.

Datasets We evaluate PEMAND on three datasets across two domains: NHTS 2017 (Federal Highway Administration, 2017) and Puget Sound 2023 (Puget Sound Regional Council, 2023) for trip planning, and PSID 2023 (Survey Research Center, 2023) for residential mobility. We use consistent preprocessing, variable selection, and train–test splits across datasets, with details in Appendix B.

Evaluation Methods. Because PEMAND covers both regression and classification tasks, we use domain-specific metrics. For trip planning on NHTS and Puget Sound, we report RMSE, MAE, sMAPE, and ±2 trip accuracy. For PSID relocation prediction, we report accuracy, F1, precision, and recall. Metric definitions are provided in Appendix B.5.

We evaluate the predictive performance of PEMAND across three datasets and compare it with both traditional ML and LLM-based baselines. The results are reported in Table 1. Across both domains, the Persona baseline consistently outperforms the Demographics-only baseline. Persona reduces MAE from 4.06 to 3.75 on NHTS and from 3.31 to 2.89 on Puget Sound, while improving F1 from 0.40 to 0.44 on PSID. This suggests that HA-CoPB-based persona modeling helps capture behavioral antecedents more effectively than raw demographic features alone. RAG and Self-Consistency do not consistently outperform

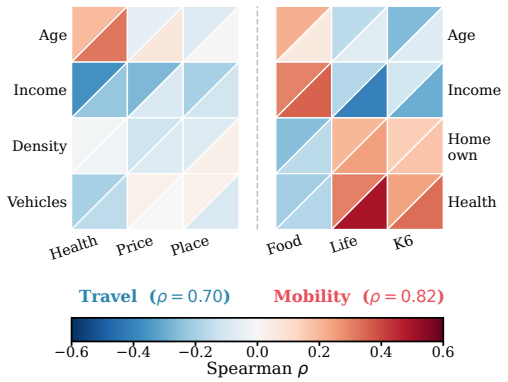

Figure 2: **Structural Correlation Alignment.** Diagonal-split heatmaps for demographic–opinion Spearman correlations in Travel (NHTS) and Mobility (PSID); upper-left triangle = human ground truth, lower-right = persona.

the single-pass Persona baseline, suggesting that generic retrieval or sampling strategies add limited value for novel household configurations. PEMAND achieves the best performance across all datasets, outperforming the strongest traditional baselines. It reduces MAE from 3.07 to 2.38 on NHTS and from 2.75 to 1.99 on Puget Sound, and improves F1 from 0.55 to 0.73 on PSID.

4.3 Perception Validation of Synthesized Personas

To assess persona coherence, we conduct a cross-domain *Perception Validation* experiment using held-out Likert-scale probes. We evaluate *Health Status*, *Price Sensitivity*, and *Walkability* in the travel domain, and *Food Affordability*, *Life Satisfaction*, and *K6 Psychological Distress* in the mobility domain. This tests whether HA-CoPB personas infer psychologically plausible perceptions from demographics rather than simply reproducing input facts.

As shown in Figure 2, HA-CoPB personas preserve latent demographic–opinion correlations in both domains, with Structural Correlation Alignment of $\rho = 0.70$ for Travel and $\rho = 0.82$ for Mobility. The figure compares human and persona Spearman correlations across key demographic variables, while individual calibration remains strong, with aggregate ± 1 accuracies of 74.2% on NHTS and 90.3% on PSID. Full probe definitions, tabular results, and metric formulations are provided in Appendix G.

4.4 Ablation Study

To assess each component, we conduct an ablation study on the Puget Sound 2023 dataset with five configurations: initial proposal without persona-alignment control, proposal with persona-based SFT, SFT-based household negotiation without moderator or ReST, negotiation with moderator control, and the full PEMAND model with ReST-based trajectory selection.

The ablation results demonstrate the contribution of each component. Without SFT, the LLM-based agents exhibit substantial persona-alignment errors, including role hallucination, violations of household constraints, and invalid trip proposals. Introducing household negotiation improves over directly aggregating individual proposals, indicating that intra-household deliberation provides useful corrective signals. However, the gain remains limited without alignment control, since hallucinated or invalid proposals can propagate through the discussion and bias the final decision. Adding intra-conversation control substantially improves performance by filtering these failure modes and stabilizing the negotiation process. Finally, incorporating ReST further improves the full model by leveraging ground-truth supervision from the training set to select higher-quality discussion trajectories for fine-tuning.

Model	MAE↓	RMSE↓	sMAPE↓	±2 Acc↑
Init. Proposal w/o SFT	3.70	5.32	63.44	0.51
Init. Proposal w/ SFT	3.10	4.47	57.90	0.56
Refinement w/o RL & Moderator	3.05	4.32	56.02	0.58
Refinement w/o RL	2.59	3.91	53.80	0.64
Consensus Refinement	**1.99**	**3.25**	**47.92**	**0.78**

Table 2: **Ablation Study on Consensus Refinement. Bold** indicates best performance. ↓ indicates lower is better, ↑ indicates higher is better.

5 Discussion - Why It Works

The model's ability to produce realistic and accurate household-level decisions stems from the careful design of each component in the proposed framework.

A key innovation is theory-guided, household-aware persona generation. Instead of using unconstrained textual personas, PEMAND grounds agents in TPB constructs and shared household context, enabling decision-relevant reasoning while preserving role and resource consistency. This turns persona generation into a structured behavioral abstraction, reducing hallucination and improving the plausibility of household decisions.

PEMAND decomposes household prediction into parallel individual proposal generation and consensus refinement. This structure reflects that household decisions emerge through interaction among members with different roles, constraints, and needs, rather than from simple aggregation. By separating initial preference elicitation from negotiation, the framework produces decisions that are both individually grounded and collectively consistent.

The framework further improves behavioral fidelity through intra- and out-of-conversation alignment control. By combining human-calibrated LLM judging, proposal-pair supervision, and ground-truth-guided trajectory selection, PEMAND reduces invalid or persona-inconsistent dialogue and refines agent policies toward more realistic household-level decisions.

6 Conclusion

We present PEMAND, a theory-guided multi-agent framework for modeling household-level decision-making with LLMs. By integrating household-aware persona generation, structured individual proposal elicitation, consensus-based refinement, and multi-level alignment control, PEMAND captures both individual heterogeneity and intra-household negotiation. Experiments across national and regional datasets show that PEMAND consistently improves predictive accuracy over traditional machine learning baselines and standard LLM-based approaches. These results suggest that structured, behaviorally grounded multi-agent LLM systems offer a promising direction for modeling collective human decisions in socially situated domains.

Limitations

This work has several limitations. First, although PEMAND improves household-level prediction for both travel demand and relocation decisions, the negotiation process is mainly validated through outcome accuracy and persona-consistency checks, leaving richer behavioral dimensions such as trip timing, mode choice, bargaining dynamics, and relocation motivations for future study. Second, while PEMAND is designed around offline open-weight models and parallelizable inference, extending multi-agent negotiation to city-, regional-, or national-level simulation would benefit from further engineering optimizations, such as adaptive stopping, selective negotiation for high-uncertainty households, and batched deployment. Finally, PEMAND should be evaluated across more diverse household structures and extended to other collective decision-making domains, such as household economic decisions, fertility, labor supply, and residential choice.

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 8–17.

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2023. Llm-deliberation: Evaluating llms with interactive multi-agent negotiation game.

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. *Advances in Neural Information Processing Systems*, 37:83548–83599.

Mariola Acosta, Margit van Wessel, Severine Van Bommel, Edidah L Ampaire, Jennifer Twyman, Laurence Jassogne, and Peter H Feindt. 2020. What does it mean to make a 'joint'decision? unpacking intrahousehold decision making in agriculture: Implications for policy and practice. *The journal of development studies*, 56(6):1210–1229.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pages 337–371. PMLR.

Jihyun Janice Ahn and Wenpeng Yin. 2025. Prompt-reverse inconsistency: Llm self-inconsistency beyond generative randomness and prompt paraphrasing. *arXiv preprint arXiv:2504.01282*.

Icek Ajzen. 1991. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211.

Barween Al Kurdi, Enass Khalil Alquqa, Shehadeh Mofleh Al-gharaibeh, Khaled Mohammad Kayed Alhyasat, Haitham M Alzoubi, Muhammad Al Shurideh, Otabeh Al-Oran, Gouher Ahmed, and Ghanim Al-Sulaiti. 2025. Determinants influencing consumer adoption of energy-efficient home appliances in jordan: An empirical analysis. *International Journal of Energy Economics and Policy*, 15(4):780–788.

Jaliya Amarasinghe Arachchige, Sara Quach, Eduardo Roca, Benjamin Liu, Alan Wee-Chung Liew, and George Earl. 2022. Understanding high-involvement product purchase through an innovative machine learning approach: A case of housing type choice. *Journal of Consumer Behaviour*, 21(5):1057–1074.

Sepehr Golrokh Amin, Devin Rhoads, Fatemeh Fakhrmoosavi, Nicholas E Lownes, and John N Ivan. 2025. Generating individual travel diaries using large language models informed by census and land-use data. *arXiv preprint arXiv:2509.09710*.

Saumya Anand, Pritikana Das, and GR Bivina. 2026. Machine learning in household trip generation modelling: A comprehensive review. *Archives of Computational Methods in Engineering*, pages 1–31.

Saumya Anand, Pritikana Das, and Bivina Geetha Rajendran. 2025. A comparative framework for trip generation prediction: A case study of raipur city, india. *National Academy Science Letters*, pages 1–6.

Terry J Atherton and Moshe Ben-Akiva. 1976. *Transferability and updating of disaggregate travel demand models*. 610.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: Llm hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156.

Nina Buchmann, Pascaline Dupas, and Roberta Ziparo. 2025. The good wife? reputation dynamics and financial decision-making inside the household. *American Economic Review*, 115(2):525–570.

Ruxiao Chen, Chenguang Wang, Yuran Sun, Xilei Zhao, and Susu Xu. 2025. From perceptions to decisions: Wildfire evacuation decision prediction with behavioral theory-informed llms. *arXiv preprint arXiv:2502.17701*.

Shaylin Chetty, Hai Le Vu, Hao Wang, and Russell Smyth. 2025. An llm framework for inferring household energy consumption through behaviour simulation. In *Proceedings of the 12th ACM International*

Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, pages 234–237.

William AV Clark and William Lisowski. 2017. Decisions to move and decisions to stay: Life course events and mobility outcomes. *Housing Studies*, 32(5):547–565.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Thomas J Cova, Yuran Sun, Xilei Zhao, Yepeng Liu, Erica D Kuligowski, Nima Janfeshanaraghi, and Ruggiero Lovreglio. 2024. Destination unknown: Examining wildfire evacuee trips using gps data. *Journal of Transport Geography*, 117:103863.

Diansheng Dong and Harry M Kaiser. 2008. Studying household purchasing and nonpurchasing behaviour for a frequently consumed commodity: two models. *Applied Economics*, 40(15):1941–1951.

Alet C Erasmus, Sune Donoghue, and Thomas Dobbelstein. 2014. Consumers perception of the complexity of selected household purchase decisions. *Journal of Retailing and Consumer Services*, 21(3):293–305.

Federal Highway Administration. 2017. 2017 national household travel survey. U.S. Department of Transportation, Washington, DC.

Yi Feng, Jiaqi Wang, Wenxuan Zhang, Zhuang Chen, Shen Yutong, Xiyao Xiao, Minlie Huang, Liping Jing, and Jian Yu. 2025. Reframe your life story: Interactive narrative therapist and innovative moment assessment with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24495–24520.

Ali Firoozzare, Flavio Boccia, Nazanin Yousefian, Sima Ghazanfari, and Somayyeh Pakook. 2024. Understanding the role of awareness and trust in consumer purchase decisions for healthy food and products. *Food Quality and Preference*, 121:105275.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Ivar Frisch and Mario Giulianelli. 2024. LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 102–111, St. Julians, Malta. Association for Computational Linguistics.

Wayne Gao, Sukjin Han, and Annie Liang. 2026. How well do llms predict human behavior? a measure of their pretrained knowledge. *Preprint*, arXiv:2601.12343.

Thomas F Golob and Michael G McNally. 1997. A model of activity participation and travel interactions between household heads. *Transportation Research Part B: Methodological*, 31(3):177–194.

Google DeepMind. 2025. google/gemma-3-27b-it. https://huggingface.co/google/gemma-3-27b-it. Hugging Face model repository.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Z. Hu, J. Lian, Z. Xiao, M. Xiong, Y. Lei, T. Wang, K. Ding, Z. Xiao, N.J. Yuan, and X. Xie. 2025. Population-aligned persona generation for llm-based social simulation. *arXiv preprint arXiv:2509.10127*.

Rob J Hyndman and George Athanasopoulos. 2018. *Forecasting: principles and practice*. OTexts.

Osman İnan and Sema Konyalı. 2025. Ethical and responsible food purchasing decisions of consumers within the scope of sustainable food policies: A case study of istanbul province. *Sustainability*, 17(11):4843.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Dezhang Kong, Shi Lin, Zhenhua Xu, Zhebo Wang, Minghao Li, Yufeng Li, Yilun Zhang, Hujin Peng, Xiang Chen, Zeyang Sha, and 1 others. 2025. A survey of llm-driven ai agent communication: Protocols, security risks, and defense countermeasures. *arXiv preprint arXiv:2506.19676*.

Erica D Kuligowski, Xilei Zhao, Ruggiero Lovreglio, Ningzhe Xu, Kaitai Yang, Aaron Westbury, Daniel Nilsson, and Nancy Brown. 2022. Modeling evacuation decisions in the 2019 kincade fire in california. *Safety science*, 146:105541.

Yu Lei, Hao Liu, Chengxing Xie, Songjia Liu, Zhiyu Yin, Canyu Chen, Guohao Li, Philip Torr, and Zhen Wu. 2024. Fairmindsim: Alignment of behavior, emotion, and belief in humans and llm agents amid ethical dilemmas. *arXiv preprint arXiv:2410.10398*.

Michael Leyer and Barbara Freytag-Leyer. 2025. Understanding the cognitive reasons why household members participate in household decision-making. *The Bottom Line*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Xuchuan Li, Fei Huang, Jianrong Lv, Zhixiong Xiao, Guolong Li, and Yang Yue. 2024. Be more real: Travel diary generation using llm agents and individual profiles. *arXiv preprint arXiv:2407.18932*.

Sung-Yoo Lim, Koki Sato, Kiyoshi Takami, Giancarlos Parady, and Eui-Jin Kim. 2026. Can large language models interpret unstructured chat data on dynamic group decision-making processes? evidence on joint destination choice. *arXiv preprint arXiv:2601.05582*.

Yifan Liu, Xishun Liao, Haoxuan Ma, Brian Yueshuai He, Chris Stanford, and Jiaqi Ma. 2025. Human mobility modeling with household coordination activities under limited information via retrieval-augmented llms. *Preprint*, arXiv:2409.17495.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46518–46535.

Xinyi Mu, Dujuan Yang, and Qi Han. 2026. Collective decision-making in energy transitions: A systematic review and a way forward. *Renewable and Sustainable Energy Reviews*, 234:116882.

Saber Naseralavi, Mohammad Soltanirad, Erfan Ranjbar, Martin Lucero, Fateme Gorzin, Yasaman Hakiminejad, Shiva Azimi, Mahdi Baghersad, and Akram Mazaheri. 2025. Machine learning modeling of household trip generation by state using nhts data. *Urban Science*, 9(9):353.

OpenAI. 2026. *GPT-5.4 mini*. Accessed: 2026-04-20.

Oude Pijp. Insights from llm-based agent model simulations. *Understanding Household Solar PV Adoption: An Innovative Approach with Large Language Models*, page 81.

Puget Sound Regional Council. 2023. 2023 puget sound household travel survey. Seattle, WA.

Baraah Qawasmeh. 2024. Estimation of a household trip-based generation model for the state of michigan. In *International conference on Environmental Design, Material Science, and Engineering Technologies*, pages 105–112. Springer.

M. Sameen, Xilei Zhao, and 1 others. 2025. Synthesizing Attitudes, Predicting Actions (SAPA): Behavioral theory-guided LLMs for ridesourcing mode choice modeling. *arXiv preprint arXiv:2509.18181*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

X. Shao and 1 others. 2024. Chain-of-planned-behavior prompting for grounded large language models. *arXiv preprint arXiv:2402.09836*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Philipp Stark, Alexandros Sopasakis, Ola Hall, and Markus Grillitsch. 2025. Using text-based life trajectories from swedish register data to predict residential mobility with pretrained transformers. *Preprint*, arXiv:2512.07865.

Yuran Sun, Shih-Kai Huang, and Xilei Zhao. 2024a. Predicting hurricane evacuation decisions with interpretable machine learning methods. *International Journal of Disaster Risk Science*, 15(1):134–148.

Yuran Sun, Ruggiero Lovreglio, Erica Kuligowski, Rosie Morrison, Thomas Cova, and Xilei Zhao. 2024b. Investigating wildfire evacuation decisions using hybrid choice modeling. *Available at SSRN*, 4953233.

Survey Research Center. 2023. Panel study of income dynamics, public use dataset. Computer file.

Mir Aftab Hussain Talpur, Shabir Hussain Khahro, Tauha Hussain Ali, Hassam Bin Waseem, and Madzlan Napiah. 2023. Computing travel impendences using trip generation regression model: a phenomenon of travel decision-making process of rural households: Mah talpur et al. *Environment, Development and Sustainability*, 25(7):5973–5996.

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

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Jiawei Wang, Renhe Jiang, Chuang Yang, Zengqing Wu, Makoto Onizuka, Ryosuke Shibasaki, Noboru Koshizuka, and Chuan Xiao. 2024. Large language models as urban residents: An llm agent framework for personal mobility generation. *Advances in Neural Information Processing Systems*, 37:124547–124574.

Peizhen Wang, Ligang Lyu, and Jiangang Xu. 2021. Factors influencing rural households' decision-making behavior on residential relocation: willingness and destination. *Land*, 10(12):1285.

Yingxu Wang and Guenther Ruhe. 2007. The cognitive process of decision making. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 1(2):73–85.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Simon Washington, Matthew G Karlaftis, Fred Mannering, and Panagiotis Anastasopoulos. 2020. *Statistical and econometric methods for transportation data analysis*. Chapman and Hall/CRC.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*.

Shenghan Wu, Yimo Zhu, Wynne Hsu, Mong-Li Lee, and Yang Deng. 2025. From personas to talks: Revisiting the impact of personas on llm-synthesized emotional support conversations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5439–5453.

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

Fei Xue and Enjian Yao. 2022. Adopting a random forest approach to model household residential relocation behavior. *Cities*, 125:103625.

Bingyu Yan, Zhibo Zhou, Litian Zhang, Lian Zhang, Ziyi Zhou, Dezhuang Miao, Zhoujun Li, Chaozhuo Li, and Xiaoming Zhang. 2025. Beyond self-talk: A communication-centric survey of llm-based multi-agent systems. *arXiv preprint arXiv:2502.14321*.

Shuo Yang, Wei Deng, Qionghua Deng, and Pengming Fu. 2016. The research on prediction models for urban family member trip generation. *KSCE Journal of Civil Engineering*, 20(7):2910–2919.

Tao Yang, Yuhua Zhu, Xiaojun Quan, Cong Liu, and Qifan Wang. 2025. Psyplay: Personality-infused role-playing conversational agents. *arXiv preprint arXiv:2502.03821*.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. 2024. Cut the crap: An economical communication pipeline for llm-based multi-agent systems. *arXiv preprint arXiv:2410.02506*.

Qin Zhang, Kelly J Clifton, Rolf Moeckel, and Jaime Orrego-Oñate. 2019. Household trip generation and the built environment: Does more density mean more trips? *Transportation research record*, 2673(5):596–606.

A Experimental Setup & Baselines

A.1 Baseline Models

To provide a comprehensive benchmark for the proposed agent-based approach, we established a set of conventional statistical and machine learning baseline models. These baselines were selected because they are widely used and cover a range of modeling assumptions and complexity, from interpretable linear models to flexible nonlinear learners.

The statistical baselines include linear regression, Poisson regression, and negative binomial regression, which are commonly applied to count-based travel outcomes. Poisson regression serves as a standard framework for modeling trip counts, while negative binomial regression is included to account for potential over-dispersion in household trip frequency. Furthermore, for the Mobility Domain (PSID), logistic regression was used to baseline the binary decision.

In addition, we evaluated several machine learning baselines, including random forest regression, gradient boosting regression, and a multi-layer perceptron (MLP) regressor/classifier, which can capture nonlinear relationships and higher-order interactions among socio-demographic, built environment, and contextual features without requiring strong distributional assumptions.

A.2 Implementation and Computational Details

A.2.1 Supervised Fine-Tuning (SFT) Hyperparameters

For the supervised fine-tuning phase, we employed QLoRA to efficiently adapt Llama-3.1-8B-Instruct while keeping the backbone frozen and quantized to 4-bit (NF4) with double quantization and bfloat16 compute dtype. LoRA adapters were attached to the attention and MLP projections (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj) using a rank of 16, an alpha of 32, and a dropout rate of 0.05. We utilized the paged 8-bit AdamW optimizer with a learning rate of 2×10^{-4}, a cosine learning-rate schedule with a 3% warmup ratio, and an effective batch size of 8 (per-device batch size of 2 with 4 gradient-accumulation steps). Training was conducted for 3 epochs with a maximum sequence length of 1024 tokens, gradient checkpointing enabled, bfloat16 mixed precision, and a fixed random seed of 0. Checkpoints were saved at the end of each epoch, retaining the two most recent. Fine-tuning was executed on a single GPU with approximately 24 GB of memory.

A.3 Parameter Setting

We set the maximum number of discussion rounds to $R_{\max} = 5$ and the per-turn revision clamp to $\rho_{\text{turn}} = 0.5$. The downward and upward bounding parameters are set to $\rho_{\text{down}} = 1.0$ and $\rho_{\text{up}} = 0.4$,

respectively. For reward weighting, we set the ground-truth term to $\omega_{GT} = 2.0$ and the rubric-based term to $\omega_R = 0.5$.

A.3.1 Inference Cost and Latency

The multi-agent framework requires several sequential steps per household, including persona proposal generation, moderator validation, and iterative discussion. Across the evaluation datasets, processing a single household required, on average, 12.72, 16.31, and 17.88 model calls for NHTS 2017, Puget Sound 2023, and PSID, respectively. The corresponding average input-token usage was 8,854.60, 13,324.90, and 19938.4, while the average output-token usage was 2,029.24, 3,561.37, and 2447.4. The average end-to-end inference runtime per household was 19.65, 22.54, and 40.02 seconds, respectively. While this is more expensive than a one-shot prediction baseline, the cost reflects the additional computational complexity required to model coordinated, constraint-aware household decision-making rather than relying on a monolithic generation.

B Dataset and Preprocessing Details

B.1 Dataset Descriptions

B.1.1 NHTS 2017

We use the household and person files from the National Household Travel Survey (NHTS) 2017 (Federal Highway Administration, 2017), a nationally representative travel survey conducted by the U.S. Federal Highway Administration. The person file contains approximately 264,235 records corresponding to individual household members, while the household file includes approximately 129,697 household-level records.

Explanatory variables are selected to capture factors known to influence household trip frequency, including socio-demographics (e.g., household composition and age structure), household attributes (e.g., income, household size, and vehicle ownership), built environment characteristics (e.g., land-use mix and density measures), and economic context variables (e.g., travel-day gas prices). Built environment and economic context variables are available only in the NHTS dataset.

Categorical variables are transformed using one-hot encoding. Since all baseline models operate at the household level, person-level attributes are aggregated into household-level features using within-household proportions (e.g., percentage of workers or students) or averages (e.g., physical activity level).

For model evaluation, we use a 90/10 train–test split, following common practice in large-scale travel behavior modeling.

B.1.2 Puget Sound 2023

To assess generalizability, we include a regional case study using the Puget Sound household travel survey (2023) (Puget Sound Regional Council, 2023). The dataset contains approximately 7,562 person records and 3,871 household records. We apply the same variable selection, feature engineering, and preprocessing pipeline as used for the NHTS dataset, excluding built environment and economic context variables that are not available in this survey.

Models are evaluated using an 80/20 train–test split, reflecting the smaller sample size of the regional dataset.

B.1.3 Panel Study of Income Dynamics (PSID) 2023

We use the 2023 family file and the 1968 - 2023 individual files from the Panel Study of Income Dynamics (PSID), a longitudinal household panel survey that has followed U.S. families and their descendants since 1968. The 2023 family file contains 9,152 household-level records, while the individual file contains 85,536 person records. The prediction task is binary classification of target variable: **residential mobility**, which means that a household changed their residence between 2021 and 2023 PSID waves. The processed household level dataset contains 9152 households, with approximately 30% of the households labeled as they moved recently. The processed PSID features are prepared by removing identifiers, redundant locations, encoding categorical variables, scaling continuous variables, and retaining ordinal and percentage based household measures in their original form.

B.2 Data Preprocessing and Imputation

B.2.1 Missing-Data Handling

Because missing and "skip-coded" values are common in survey data, we applied multiple strategies to handle incomplete records across all datasets:

1. **Household-level removal:** If any household member contained invalid or missing values for critical variables, the entire house-

hold was excluded to maintain consistency in household-level aggregation.

2. **Rule-based recoding:** Selected missing or negative-coded responses were recoded using domain-informed rules. For example, negative codes for rideshare usage were treated as no rideshare use.

3. **Median/mode imputation:** For remaining variables with missing values, numeric features were imputed using the median and categorical features were imputed using the mode to reduce data loss while preserving overall distributions.

B.2.2 Feature Engineering and Scaling

For the classical machine learning baselines, datasets were preprocessed using the following steps:

- **Categorical Encoding:** Categorical attributes (excluding those already ordinal encoded) were transformed using one-hot encoding.

- **Aggregation (Travel Domain):** Since all baseline models operate at the household level, person-level attributes in the NHTS and Puget Sound datasets were aggregated into household-level features using within-household proportions (e.g., percentage of workers or students) or averages (e.g., physical activity level).

- **Scaling (Mobility Domain):** For the PSID dataset, identifiers and redundant locations were removed. Percentage features were retained in their original form. The remaining numerical features were first power-transformed using the Yeo-Johnson method before standardization to handle skewness.

B.3 Variable Definitions

Table 11 lists the variables used in the travel domain experiments. Variables marked with "*" are included in both the NHTS and Puget Sound datasets. Table 12 lists the variables utilized for the processed PSID mobility dataset.

B.4 Coding Schemes

The coding conventions used for binary, frequency-based, and income bracket variables are summarized in Table 3.

B.5 Evaluation Metrics Details

We compare predicted household trip counts with observed values using multiple complementary evaluation metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), symmetric Mean Absolute Percentage Error (sMAPE), and a tolerance-based accuracy measure. Together, these metrics capture overall goodness-of-fit, absolute and relative error magnitudes, and the practical usefulness of predictions for household-level travel decision-making and transportation planning contexts.

Furthermore, we used accuracy, precision, recall, and F1 score for binary classification evaluations.

Root Mean Squared Error (RMSE) RMSE measures the square root of the average squared difference between predicted and actual values:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2}. \tag{8}$$

By squaring the errors, RMSE penalizes large deviations more heavily and is therefore sensitive to extreme prediction errors.

Mean Absolute Error (MAE) MAE computes the average absolute difference between predicted and actual values:

$$\text{MAE} = \frac{1}{n}\sum_i |y_i - \hat{y}_i|. \tag{9}$$

Compared to RMSE, MAE is less influenced by outliers and provides an interpretable measure of the typical prediction error in the original unit of trips.

Symmetric Mean Absolute Percentage Error (sMAPE) sMAPE is a scale-independent metric that evaluates relative prediction error:

$$\text{sMAPE} = \frac{1}{n}\sum_i \frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}. \tag{10}$$

Household trip counts exhibit a high frequency of zero-valued observations, for which conventional percentage-based error metrics such as MAPE become undefined or numerically unstable. By symmetrically normalizing prediction errors using both predicted and actual values, sMAPE provides a more stable evaluation of relative error in zero-inflated travel demand data.

Variable Type	Code / Value	Meaning
Binary	0	No
Binary	1	Yes
Frequency	01	Never
Frequency	02	A few times a year
Frequency	03	A few times a month
Frequency	04	A few times a week
Frequency	05	Daily
Income Bracket	1	Less than $25k
Income Bracket	2	$25k – $50k
Income Bracket	3	$50k – $75k
Income Bracket	4	$75k – $100k
Income Bracket	5	Higher than $100k

Table 3: Coding schemes for binary, frequency, and income bracket variables.

Accuracy within a Tolerance (± 2 Trips) To assess practical prediction performance, we report a tolerance-based accuracy metric defined as the proportion of households whose predicted trip counts fall within ± 2 trips of the actual values. Formally, a prediction for household i is considered accurate if:

$$|y_i - \hat{y}_i| \leq 2, \tag{11}$$

and the reported accuracy is computed as the fraction of households satisfying this condition across the evaluation set. Household-level trip counts are obtained by aggregating discrete individual trip counts across household members, most of whom form households of two or more individuals.

As a result, small deviations at the individual level naturally translate into differences of one to two trips at the household level, making a ± 2 tolerance a reasonable margin of error given the data aggregation process.

Accuracy Accuracy measures the proportion of correctly classified instances among all evaluated cases:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{12}$$

Accuracy provides an intuitive summary of overall classification performance but can be misleading under class imbalance, where a trivial majority-class predictor may still attain a high score.

Precision Precision quantifies the proportion of predicted positives that are truly positive:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{13}$$

Precision is informative when false positives are costly, as it reflects how reliable a positive prediction is.

Recall Recall measures the proportion of actual positives that are correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{14}$$

Recall is particularly relevant when false negatives carry a high cost, as it captures how completely the model retrieves positive cases.

F1 Score The F1 score is the harmonic mean of precision and recall, balancing the two into a single summary statistic:

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{15}$$

By combining precision and recall, the F1 score is particularly useful under class imbalance, where accuracy alone may overstate performance.

C Additional Main Results

C.1 Robustness Across Different LLM Backbones

To evaluate the robustness of our framework across different model architectures, we conducted a side-by-side comparison deploying homogeneous pipelines end-to-end using Llama-3.1-8B-Instruct, Gemma-3-27b-it, and GPT-5.4-mini on the Puget Sound dataset (Table 4). The results demonstrate that PEMAND consistently outperforms all single-agent baseline variants regardless of the underlying LLM backbone. We observe several key insights:

First, performance across zero-shot single-agent baselines are highly model-dependent. While the Demographics-only baseline struggles on Llama (MAE 3.31), it performs well on Gemma (MAE 2.83, ± 2 accuracy 0.58), surpassing both Llama and GPT-5.4-mini. Second, single-agent persona conditioning is not uniformly beneficial. While injecting a behavioral persona improves Llama's

Method	Llama-3.1-8B-Instruct				Gemma-3-27b-it				GPT-5_4-mini			
	MAE↓	RMSE↓	sMAPE↓	±2 Acc↑	MAE↓	RMSE↓	sMAPE↓	±2 Acc↑	MAE↓	RMSE↓	sMAPE↓	±2 Acc↑
Demographics-only	3.31	5.10	61.75	0.54	2.83	4.19	55.63	0.58	2.98	4.23	56.47	0.55
Persona	2.89	3.99	59.52	0.56	3.33	5.37	62.99	0.55	3.09	4.25	56.13	0.53
RAG	5.01	6.60	76.49	0.35	3.93	5.43	66.85	0.43	3.46	4.71	60.61	0.47
Self-consistency	2.89	4.04	58.30	0.57	3.01	4.27	61.70	0.55	2.91	4.07	54.34	0.57
PEMAND w/o ReST	2.59	3.91	53.80	0.64	2.41	3.64	49.54	0.68	**2.32**	**3.78**	**47.87**	**0.63**
PEMAND w ReST	**1.99**	**3.25**	**47.92**	**0.78**	**1.68**	**2.98**	**45.73**	**0.81**	—	—	—	—

Note: We do not report **PEMAND w ReST** for GPT-5.4-mini because QLoRA adapter training was applied only to open-weight backbones; GPT results use API inference without weight updates.

Table 4: **LLM Baselines on Puget Sound 2023 (test).** Comparison of **Llama-3.1-8B-Instruct**, **Gemma-3-27b-it**, and **GPT-5_4-mini** under the same baseline families, with a *homogeneous* backbone in each column (persona generation, member agents, moderator, and judge share the same model family). **Bold** is best among the four LLM baselines within each model block and metric column. **PEMAND w/o ReST / PEMAND w ReST**: multi-agent pipeline without or with QLoRA SFT on member models for that backbone column. ↓ / ↑ indicate lower / higher is better.

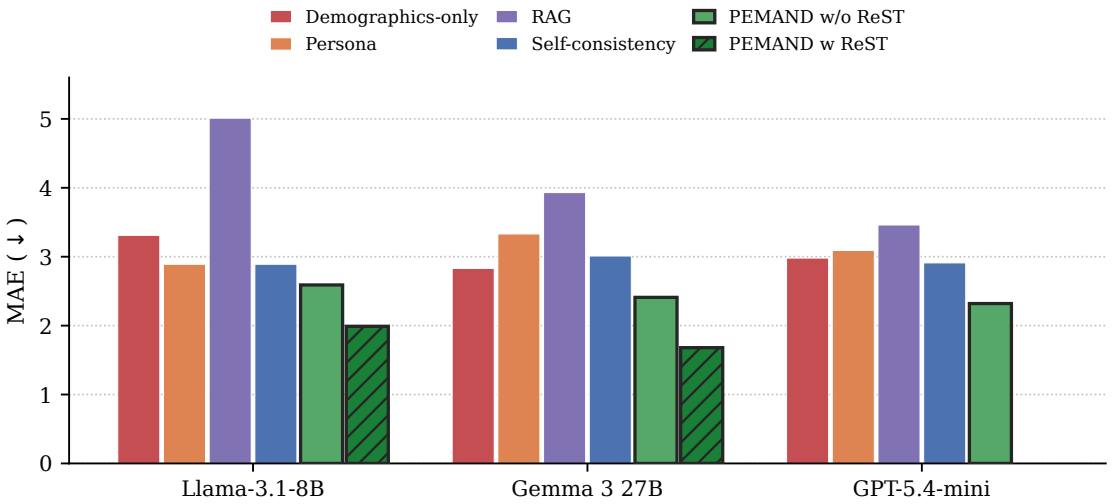

Figure 3: **Robustness across homogeneous LLM backbones (Puget Sound test, MAE).** Grouped bars compare baseline families under Llama-3.1-8B-Instruct, Gemma-3-27b-it, and GPT-5.4-mini. PEMAND achieves the lowest MAE on every backbone; stronger models reduce error for most families, though persona conditioning is not uniformly beneficial. PEMAND w ReST is omitted for GPT-5.4-mini (QLoRA training applied only to open-weight backbones; Table 4).

prediction (reducing MAE to 2.89), it degrades performance for both Gemma (MAE 3.33) and GPT-5.4-mini (MAE 3.09) compared to their respective Demographics-only baselines. Retrieval-Augmented Generation (RAG) remains universally harmful across all three backbones, likely because retrieving historical households with conflicting trip counts induces severe covariate shift. By contrast, Self-consistency remains a strong single-agent strategy, particularly for GPT-5.4-mini (MAE 2.91) and Gemma (MAE 3.01), by reducing sampling variance.

Ultimately, the complete PEMAND multi-agent negotiation framework achieves the best overall performance across backbone configurations. The Gemma-3-27B model yields stronger overall results than Llama-3.1-8B, which is expected given its larger parameter scale and stronger generation capacity. Importantly, however, PEMAND remains highly competitive even with the lighter Llama backbone, indicating that the performance gains do not rely solely on model size. Instead, the results suggest that the structured negotiation protocol, alignment control, and trajectory refinement provide robust error reduction across different LLM backbones. This supports the generality of the proposed framework and shows that PEMAND can benefit from larger frontier-scale models while remaining effective under more lightweight and deployable open-weight settings.

D Prompt Engineering Interfaces

We formalize our prompt designs as structured cognitive interfaces. Each template employs explicit role definition, context delimitation, and negative constraint protocols to ensure reproducibility and mitigate LLM hallucination.

Domain Generalization Mapping: To support multiple decision domains, our prompt architectures use unified structural placeholders that are dynamically instantiated based on the target dataset. Table 5 illustrates the mappings across the two primary case studies.

D.1 Phase 1: Faithfulness-Constrained Narrative Synthesis

The following prompts drive the transformation of tabular encoded data into natural language personas. We employ a Clinical Neutrality Protocol to strip RLHF-induced positivity bias. To prevent the persona from embedding target decision outcomes, we impose task-specific vocabulary restrictions. For example, in the residential mobility task, we explicitly ban evaluative adjectives (e.g., "stable", "rooted", "transient") and predictive language (e.g., "likely to move", "intends to relocate") to avoid outcome-leakage or stereotyping. We also enforce relationship-aware constraints to ensure syntactical consistency with the individual's role relative to the household head.

Box 1: Persona Generation System Definition

[ROLE DEFINITION]
You are a factual data-to-text transcription engine. Your objective is to convert atomic variable states into a coherent first-person narrative without semantic alteration or emotional embellishment.
[STRICT CONSTRAINTS]
1. **Fidelity**: Use every provided fact exactly as given. Do not hallucinate traits not present in the input.
2. **Sentiment Neutralization**: The following subjective descriptors are **BANNED**: proud, blessed, fortunate, thrilled, excited, grateful.
3. **Role Rigor**: Do not infer authority phrases like "Head of Household" or "responsible for" unless explicitly encoded in the relationship variable.
4. **Vocabulary Restrictions**: stable, rooted, transient, likely to move, intends to relocate are **BANNED**.
5. **Scope**: NEVER mention the target variable. Output ONLY the persona text.

Box 2: Persona Generation Input Interface

[INPUT DATA: ATOMIC FACTS]
The following list represents the ground-truth state of the agent:
{facts}
[GENERATION INSTRUCTIONS]
Synthesize a first-person narrative adhering to the following logic:
- **Identity**: Begin with "I am..."
- **Completeness**: Integrate 100% of the atomic facts listed above.
- **Tone**: Maintain a clinical, objective tone. Avoid filler words.
- **Consistency**: If facts imply conflicting states (e.g., high income but zero car), state them clearly without attempting to "fix" the contradiction.
[OUTPUT]
Persona:

D.2 Phase 2: Individual Reasoning & Alignment via Refinement (AvR)

These templates drive the core HA-CoPB reasoning engine and the dual-agent alignment layer. The Actor generates the TPB evaluation, and the Reviewer calibrates it against the statistical prior.

Box 3: HA-CoPB Actor Reasoning Template

[SYSTEM CONTEXT]
You are a behavioral reasoning engine. Your goal is to predict the <target decision> for a specific household member, but you must account for the shared constraints and obligations of the entire household.
[HOUSEHOLD STATE]
Context: Size: {hh_size} | Vehicles: {veh_count} | Income: {income}
Shared Constraints: <shared constraints>
Behavioral Anchor: Similar individuals typically exhibit **{anchor}** behavior.
[PERSONA PROFILE]
Persona: {persona}
Behavioral Tendency: {marker_statement} (Implication: {marker_implication})
[HA-CoPB REASONING TASK]
Perform a Theory of Planned Behavior (TPB) analysis that entangles individual desires with household reality:
1. Attitude (Individual Utility):
- How does the "Behavioral Tendency" shift their intrinsic demand?
2. Subjective Norms (Household Obligations):
- What role-based obligations exist?
3. Perceived Behavioral Control (Resource Competition):
- Do financial, spatial, or shared resource constraints limit their autonomy?
[OUTPUT]
Rationale: <Step-by-step TPB analysis resolving the conflicts above>

Placeholder	Travel Domain (NHTS/Puget)	Mobility Domain (PSID)
<target decision>	Total daily trip frequency	Relocated between 2021–2023
<shared constraints>	Vehicle deficit/bottlenecks	Homeownership lock-in
<agent role>	Travel Diary Participant	Head of Household/Spouse
<output format>	Integer (e.g., 4)	Boolean (True/False)

Table 5: Dynamic prompt instantiation across domains.

Box 4: AvR Moderator Calibration Template

[SYSTEM CONTEXT]
You are the Lead Demographic Sociologist and Calibration Moderator. Your task is to review the Theory of Planned Behavior reasoning generated by the primary behavioral analyst and issue the final calibrated individual intention.
[EMPIRICAL BASELINE (ANCHOR)]
Households with demographics similar to this one have an empirical baseline rate of **{anchor}**.
[CALIBRATION RULES]
1. Treat this empirical anchor as a qualitative prior expectation, NOT a strict mathematical threshold.
2. Read the primary analyst's reasoning. If the analyst identifies overwhelming desires that overpower the macro-trend, you MUST deviate from a low baseline.
3. If the analyst identifies absolute structural vetoes (like <shared constraints>), you MUST enforce the constraint, regardless of their positive attitude.
[OUTPUT]
Provide a brief justification evaluating if the analyst's identified constraints or facilitators warrant the action, then output:
Final Answer: <output format>

D.3 Phase 3: Multi-Agent Parallel Proposal

These templates illustrate how member agents express their initial intentions and needs during the proposal phase.

Box 5: Parallel Proposal Prompt Template

[SYSTEM CONTEXT]
Role-play as one household member. Stay in character using the given sociodemographic profile and TPB perceptions. Respect household constraints and do not fabricate members or resources.
[AGENT PERSONA]
<sociodemographic profile>
[TPB PERCEPTIONS]
<TPB perceptions>
[TASK]
Independently propose your initial <decision>. Explain the reasoning in 2-4 persona-grounded sentences, then output the decision on the last line.
[OUTPUT FORMAT]
Respond **STRICTLY** as:
REASON: <2-4 sentence justification>
FINAL_<DECISION>: <value>

D.4 Phase 4: Multi-Agent Consensus Refinement

These templates illustrate how member agents interact, revise their proposals in response to other members' needs and opinions, and move toward a shared household decision.

Box 6: Consensus Refinement

[SYSTEM CONTEXT]
You are one household member discussing a shared household-level <decision>. Stay in character and revise your view when another member provides valid evidence.
[YOUR IDENTITY]
<persona + TPB perceptions + initial proposal>
[OTHER MEMBERS]
<other members' personas + initial proposals>
[HOUSEHOLD CONSTRAINTS]
<household size, income, vehicles, drivers, and non-speaking dependents>
[STARTING POINT]
Initial values: <list>
Aggregated initial <decision>: <$\hat{y}_0$>
[CURRENT VALUES]
<latest PREFERRED_<DECISION> from each speaker>
[CONVERSATION HISTORY]
<h_t – truncated dialogue history>
[TASK]
Reassess all members' evidence and either propose a revised household <decision> or use SKIP after your first turn. Acknowledge a specific point from another member, justify any change from your previous value, and respect household constraints.
[OUTPUT FORMAT]
Respond **STRICTLY** as:
ACKNOWLEDGE: <specific point from another member or initial value>
SUPPORT: <evidence-grounded justification>
PREFERRED_<DECISION>: <value>
DELTA_FROM_LAST: <brief reason for change or no change>
Or after the first turn: SKIP.

D.5 Phase 5: Discussion Moderator

These templates show how the moderator evaluates each member agent's utterance.

Box 8: Moderator Validation Prompt Template

[SYSTEM CONTEXT]
You are the moderator agent supervising a household decision-making discussion. Your role is to evaluate whether a candidate utterance satisfies the provided annotation guideline. You do **not** participate as a household member and do **not** revise the utterance.
[ANNOTATION GUIDELINE]
<annotation guideline>
[HOUSEHOLD CONTEXT]
<household context>
[SPEAKER PROFILE]
<agent persona profile>
[HOUSEHOLD PROPOSAL]
Current household decision estimate: <$\hat{y}_t$>
[CONVERSATION HISTORY]
<h_t>
[CANDIDATE UTTERANCE]
<$u_{candidate}$>
[TASK: MODERATOR VALIDATION]
Evaluate the candidate utterance according to the provided annotation guideline. Assign a score from 1 to 3 for each rubric dimension, where 3 indicates full compliance with the guideline. The candidate utterance passes validation only if it receives a score of 3 on every dimension. If any dimension receives a score below 3, the utterance must be rejected and regenerated. Rejected utterances are discarded and **not** added to the conversation history.
[OUTPUT FORMAT]
Respond with:
SCORES: <dimension_1>=<1-3>; <dimension_2>=<1-3>; ...
DECISION: PASS or DECISION: REJECT
FEEDBACK: <one short reason if rejected; otherwise "All dimensions receive full scores.">

D.6 Phase 6: Comparison Baselines

D.6.1 Baseline: Demographics Only

Box 8: Demographics-Only Baseline Interface

[SYSTEM ROLE]
You function as a direct feature-to-label predictor.
[INPUT FEATURES]
Demographic Profile:
{demographics}
[PREDICTION TASK]
Based strictly on the features provided above, predict the <target decision> for this agent.
[OUTPUT]
Provide a single-sentence rationale followed by the prediction.
Rationale: <Brief justification>
Final Answer: <output format>

D.6.2 Baseline: Persona

Box 9: Persona Baseline

[SYSTEM CONTEXT]
You are an expert Behavioral Psychologist predicting **total household behavior**.
[MEMBER PROFILES]
The following narratives describe the household members:
{member_narratives}
[PREDICTION TASK]
Predict the <target decision> using Theory of Planned Behavior. You must mentally simulate the negotiation of shared constraints:
1. Aggregated Needs (Attitude):
- Sum the mandatory needs for all members.
2. Shared Constraints (PBC):
- Bottleneck Check: Evaluate shared resource limitations.
3. Synthesis:
- Start with the Anchor ({anchor}).
- Adjust based on the net balance of Needs vs. Constraints.

[OUTPUT]
Respond STRICTLY in this format:
Needs Analysis: <Who needs the action?>
Constraint Logic: <How do limits reduce the probability?>
Final Answer: <output format>

D.6.3 Baseline: Self-Consistency

The Self-Consistency baseline does not employ a distinct prompt structure. Instead, it utilizes the *Persona* template (Box 9), but executes the inference step $K = 10$ times at a non-zero temperature ($T = 0.7$). The final prediction is generated by taking the majority vote (for classification tasks) or the median (for regression tasks) of the K sampled outputs. This baseline tests whether stochastic exploration of reasoning paths improves zero-shot predictive alignment without requiring explicit multi-agent iteration.

D.6.4 Baseline: Retrieval-Augmented Generation (RAG)

The RAG baseline augments the Household CoPB prompt by injecting $N = 3$ historically retrieved household exemplars (nearest neighbors in the training set based on demographic feature similarity) to provide local context.

Box 10: RAG Baseline Template

```
[SYSTEM CONTEXT]
You are an expert Behavioral Psychologist
predicting total household behavior.
[HISTORICAL EXEMPLARS (RETRIEVED CONTEXT)]
The following similar households from the
historical database made the following
decisions:
- Exemplar 1: {retrieved_hh_1_features} →
Outcome: {retrieved_outcome_1}
- Exemplar 2: {retrieved_hh_2_features} →
Outcome: {retrieved_outcome_2}
- Exemplar 3: {retrieved_hh_3_features} →
Outcome: {retrieved_outcome_3}

[CURRENT HOUSEHOLD: MEMBER PROFILES]
The following narratives describe the
current household members:
{member_narratives}
[PREDICTION TASK]
Predict the <target decision> using Theory
of Planned Behavior. Leverage the
historical exemplars above as context
for how similar households resolve their
constraints.
[OUTPUT]
Respond STRICTLY in this format:
Needs Analysis: <Who needs the action?>
Constraint Logic: <How do limits reduce
the probability?>
Final Answer: <output format>
```

E Behavioral Theory Implementation

This section details the formal extraction and reasoning rules used to map raw survey data into theory-grounded cognitive representations for both the Travel Behavior (NHTS/Puget Sound) and Residential Mobility (PSID) domains.

E.1 HA-CoPB Framework Overview

Figure 4 illustrates how our Household-Aware CoPB (HA-CoPB) framework adapts the classical Theory of Planned Behavior (TPB) (Ajzen, 1991) to the household decision-making setting. While the standard TPB models behavioral intention as a function of Attitude, Subjective Norms, and Perceived Behavioral Control at the individual level, our HA-CoPB reinterprets these constructs to capture household-level dynamics: $\mathcal{PBC}$ captures shared resource bottlenecks, $\mathcal{SN}$ captures concrete intra-household role obligations, and $\mathcal{A}$ captures individual intrinsic preferences.

E.2 Translation Map and Neutralization Filter Details

To ensure the persona narratives are factually grounded and free of dataset-specific structural leakage, we map raw survey codes to deterministic natural language statements. Table 6 provides the mapping logic for the travel domain, while Table 7 provides the longitudinal mapping logic for the PSID residential mobility domain.

E.3 Behavioral Anchor Logic

Consistent with the principle of temporal model transferability (Atherton and Ben-Akiva, 1976), the model treats the historical anchor as a prompt-based prior.

- **Continuous Formulation (Travel):** The prior $y_{\text{prior}}^{t-\text{lag}}$ is represented as a continuous average frequency (e.g., mean daily trip rate λ) calculated across historical demographic subgroups based on primary identifiers (e.g., age bracket, vehicle ownership, urban/rural status).

- **Probability Formulation (Mobility):** The prior $y_{\text{prior}}^{t-\text{lag}}$ is represented as a baseline relocation probability synthesized through multi-factor weighted blending of historical subpopulations. It serves as an expected base rate which the AvR Moderator evaluates against qualitative deviations in the agent's HA-CoPB reasoning.

E.4 Attitudinal Markers and Imputation Rules

The Attitude ($\mathcal{A}$) construct provides the intrinsic psychological drivers for the persona. Table 8 details the specific latent psychographic markers inferred for both the Travel and Mobility domains, alongside their specific HA-CoPB implications.

E.5 Hierarchical Priority & Construct Logic

To prevent LLM hallucination and ground predictions in behavioral realism, the generation process strictly enforces the Hierarchical Priority Mechanism ($\mathcal{PBC} \succ \mathcal{SN} \succ \mathcal{A}$).

- **Travel Domain Constraints:** An agent must evaluate hard resource limitations ($\mathcal{PBC}$) before considering normative obligations ($\mathcal{SN}$) or intrinsic desires ($\mathcal{A}$). For instance, if a household has 0 vehicles (PBC), a "Pro-Car" attitude (A) cannot trigger a private vehicle trip. Similarly, if there is a vehicle bottleneck (e.g., 2 drivers, 1 car), the model must simulate negotiation rather than hallucinating independent parallel trips.

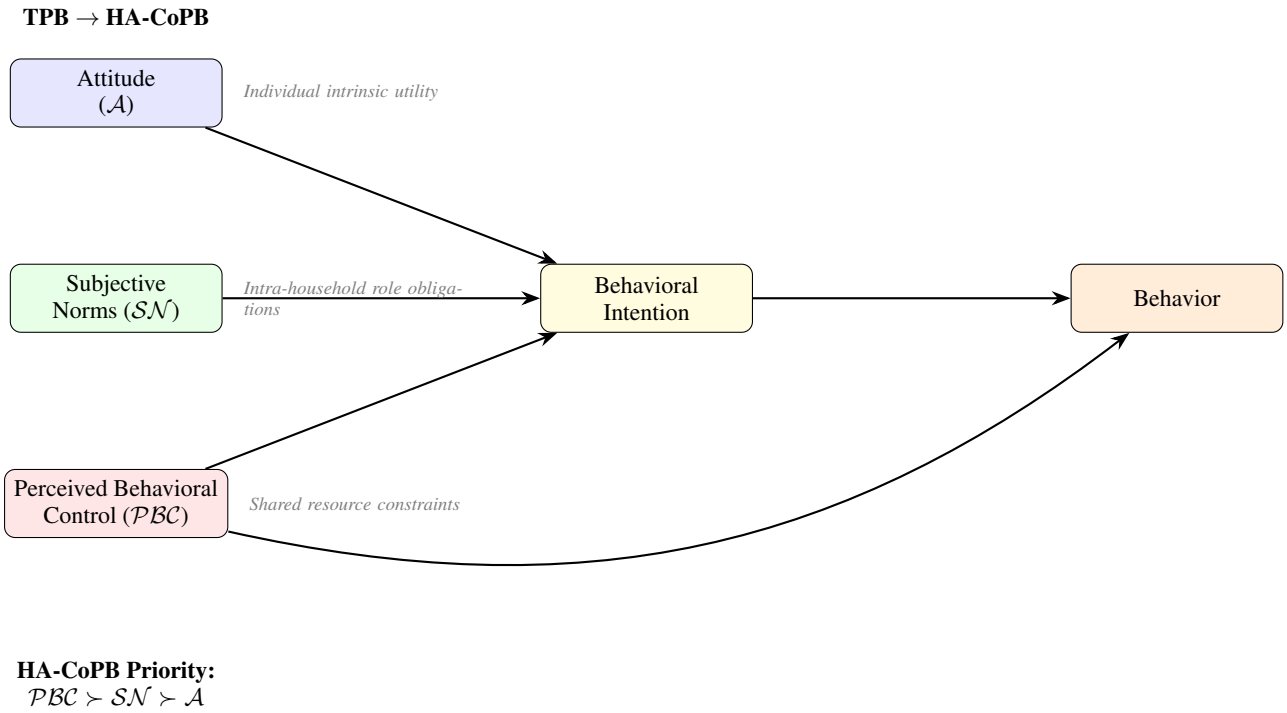

Figure 4: The Theory of Planned Behavior (TPB) and the proposed HA-CoPB adaptation. Our hierarchy mandates that shared resource barriers ($\mathcal{PBC}$) override social obligations ($\mathcal{SN}$), which in turn override individual preferences ($\mathcal{A}$).

- **Mobility Domain Constraints (PSID):** Structural immobility ($\mathcal{PBC}$) dominates desire-driven relocation ($\mathcal{A}$). If a household owns their home (a massive transaction friction) or suffers from acute labor shocks with no wealth cushion, the PBC acts as a hard veto against moving. Conversely, deep poverty often traps households structurally, overriding a desire for better neighborhoods. Only when PBC constraints are satisfied can the Subjective Norms (e.g., family crowding) and Attitudes (e.g., Career-Driven) dictate the final AvR consensus.

F Revision Clamping in Consensus Refinement

During the consensus refinement process, we did revision clamping for each household member's utterance at each round for numerical decisions (with binary decisions unchanged). Specifically, the per-round decision update is limited to at most a fraction ρ_{turn} of the speaker's previous value. In addition, the preferred decision is bounded relative to the aggregated initial proposal:

$$\hat{y} \in \left[(1-\rho_{\text{down}})\hat{y}^{(0)}, (1+\rho_{\text{up}})\hat{y}^{(0)}\right], \qquad (16)$$

where $\hat{y}$ denotes the current preferred decision, and ρ_{down} and ρ_{up} control the allowable downward and upward deviation from the initial household proposal. This bound prevents the refined decision from drifting excessively from members' initial intentions while still allowing reasonable updates during consensus refinement.

G Perception Validation Details

Survey Instruments To rigorously evaluate whether the synthesized personas possess a coherent worldview consistent with their demographics, we validate their responses against specific attitudinal variables held out from the generative phase. For the Travel Domain (NHTS), we preserve the exact wording and Likert scales used in the original survey:

- **Health Status (HEALTH):**
  Question: "In general, would you say your health is?"
  Scale: 1=Excellent, 2=Very Good, 3=Good, 4=Fair, 5=Poor.
- **Price Sensitivity (PRICE):**
  Statement: "The price of gasoline affects the number of vehicle trips I make."
  Scale: 1=Strongly Agree, 2=Agree, 3=Neither

Variable	Raw Code Example	Natural Language Mapping
Demographics		
R_AGE_IMP	34	"I am 34 years old."
R_SEX_IMP	1 (Male), 2 (Female)	"I am a man/woman." (uses "boy/girl" if Age < 18)
EDUC	5 (Grad Degree)	"I have a graduate or professional degree."
HHFAMINC	11 ($> \$200k$)	"My household income is $200,000 or more."
Household		
R_RELAT	3 (Child)	"I am the son/daughter of the household head."
HHVEHCNT	0	"My household does not own any vehicles."
LIF_CYC	4 (2+ Adults, Young Child)	"My household consists of 2 adults, 1 child, youngest child aged 0-5."
Context		
URBRUR	1 (Urban)	"I live in an urban area."
RAIL	1 (Has Rail)	"My metropolitan area has heavy rail."
PHYACT	3 (Vigorous)	"I engage in some vigorous physical activities."

Table 6: Travel Domain Variable Translation Logic. Raw codes from the travel survey are deterministically mapped to natural language atoms before narrative synthesis.

Variable	Raw Code Example	Natural Language Mapping
Demographics		
AGE	45	"I am a 45-year-old man/woman."
EMPLOYMENT_STATUS	2 (Temporarily laid off)	"Regarding employment, I am temporarily laid off."
DEGREE_TYPE	2 (Bachelor's)	"My highest degree is a bachelor's degree."
Household		
RELATION_TO_HEAD	10 (Reference Person)	"I am the reference person (head) of this household."
income_to_needs	3.5	"Household context: family income relative to census needs standard is about 3.50."
HHVEHCNT	2	"Household context: the household owns 2 vehicle(s)."
Longitudinal Context		
bh_tenure_years_2021	4.0 (Years in home)	"Prior-wave snapshot: the household's tenure at the current residence AS OF MID-2021 was at least 4 years."
bh_move_count_prior	2 (Moves in prior waves)	"Panel history: the household moved twice across those two prior waves."

Table 7: Residential Mobility (PSID) Variable Translation Logic. To avoid leaking the move-outcome, we explicitly enforce temporal labeling (e.g., "Prior-wave snapshot") and neutralize relationship categories.

Agree or Disagree, 4=Disagree, 5=Strongly Disagree.

- **Walkability Preference (PLACE):**
  Statement: "I prefer to live in a community with mixed land uses (homes, shops, work) so I can walk to places."
  Scale: 1=Strongly Agree, 2=Agree, 3=Neither Agree or Disagree, 4=Disagree, 5=Strongly Disagree.

For the Mobility Domain (PSID), we leverage longitudinal well-being and distress indicators:

- **Food Affordability (FOOD_AFFORD):**
  Statement: "We couldn't afford to eat balanced meals."
  Scale: 1=Often True, 2=Sometimes True, 3=Never True.
- **Life Satisfaction (LIFE_SAT):**
  Question: "Please think about your life-as-a-whole. How satisfied are you with it?"
  Scale: 1=Completely Satisfied to 5=Not at all Satisfied.
- **Psychological Distress (K6_DISTRESS):**
  Measure: Aggregate K6 distress scale mapped to ordinal quintiles.
  Scale: 1=Low Distress to 5=High Distress.

Tabular Results Figure 2 in the main text visualizes the diagonal-split correlation heatmaps underlying the aggregate alignment scores below. Table 9 provides the detailed quantitative breakdown of perception validation performance across all evaluated metrics for both datasets.

Distributional Histograms Figure 5 compares the marginal response distributions for all six hold-

Domain	Construct	Narrative Statement	HA-CoPB Implication
Travel	Tech-Savvy	"Learning new technologies is often frustrating for me. (Reverse)"	High PBC for app-based mobility and e-shopping substitution.
Travel	Polychronic	"I prefer to do one thing at a time. (Reverse)"	Tolerance for longer transit/ride-hail trips if productive; justifies complex trip chaining.
Travel	Materialistic	"I enjoy having a lot of luxury things."	Vehicle Preference: Correlates with private vehicle usage and higher discretionary shopping generation.
Travel	Pro-Car	"I like the idea of driving as a means of travel."	Resistance to mode switching; high utility for driving.
Travel	Pro-Transit	"I like the idea of public transit as a means of travel."	Increases probability of transit usage if PBC allows.
Travel	Wait Tolerant	"The trip itself is a waste of time. (Reverse)"	Lower disutility of travel time; longer commute acceptance.
Mobility (PSID)	Place-Attached	"I prefer to stay in my current neighborhood."	Lower baseline mobility; tenure and local ties increase moving costs in the PBC layer.
Mobility (PSID)	Career-Driven	"I am willing to relocate to advance my career."	Job-linked mobility; strengthens move intentions when employment shocks appear.
Mobility (PSID)	Family-Oriented	"Schooling and space needs come first."	Crowding increases subjective-norm pressure toward larger housing.
Mobility (PSID)	Cost-Constrained	"Housing cost burdens make it hard to move."	PBC frictions: affordability limits feasible moves; may force distress moves.
Mobility (PSID)	Lifestyle-Seeker	"I value flexibility and am open to new cities."	Higher discretionary mobility among renters and young adults.
Mobility (PSID)	Stability-Seeker	"I prioritize predictable housing and avoiding disruption."	Strong inertia; homeownership and lifecycle stability reduce realized mobility.

Table 8: Attitudinal Constructs mapped to their HA-CoPB theoretical implications across both domains.

	Travel Domain (NHTS)			Mobility Domain (PSID)		
Metric	**Health**	**Price**	**Place**	**Food Afford.**	**Life Sat.**	**K6 Distress**
Accuracy (Exact)	38.1%	26.2%	27.7%	71.0%	41.4%	47.9%
Accuracy (±1)	**85.0%**	**64.3%**	**73.4%**	**94.7%**	**88.4%**	**87.7%**
MAE	0.80	1.32	1.07	0.34	0.71	0.69
QWK (Kappa)	0.29	0.02	0.03	0.11	0.18	0.18
Wasserstein (W_1)	0.47	0.86	0.28	0.12	0.26	0.33
Spearman (r_s)	0.29	0.01	0.03	0.15	0.18	0.19
Aggregate Structural Align (ρ)		**0.70**			**0.82**	

Table 9: **Comprehensive Perception Validation Results.** Performance of PEMAND across all targeted perception probes in both domains.

out probes. Each panel overlays the human ground-truth density (dark) against the synthesized persona density (green). K6 distress is shown on quintile-binned ordinal scores (1=low to 5=high), consistent with the individual-level metrics in Table 9. Visual inspection confirms that personas track the dominant modes for well-structured probes (e.g., *Health Status*, *Food Affordability*) while revealing expected smoothing on highly subjective items (e.g., *Price Sensitivity*).

Mathematical Formulations of Metrics

Accuracy within a Tolerance (±1 Scale Point) We adapt tolerance-based accuracy to Likert-scale prediction by treating a response as accurate when $|y_{i,v} - \hat{y}_{i,v}| \leq 1$. This ±1 tolerance reflects the subjectivity of attitude ratings (e.g., the subtle distinction between "Agree" and "Strongly Agree") and captures whether the persona's sentiment is directionally correct.

Quadratic Weighted Kappa (QWK) To measure individual-level agreement while accounting for the ordinal nature of Likert scales, we use Quadratic Weighted Kappa (Cohen, 1968). Unlike simple accuracy, QWK penalizes large disagreements (e.g., predicting 1 vs 5) significantly more than minor shifts (e.g., 1 vs 2):

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}} \tag{17}$$

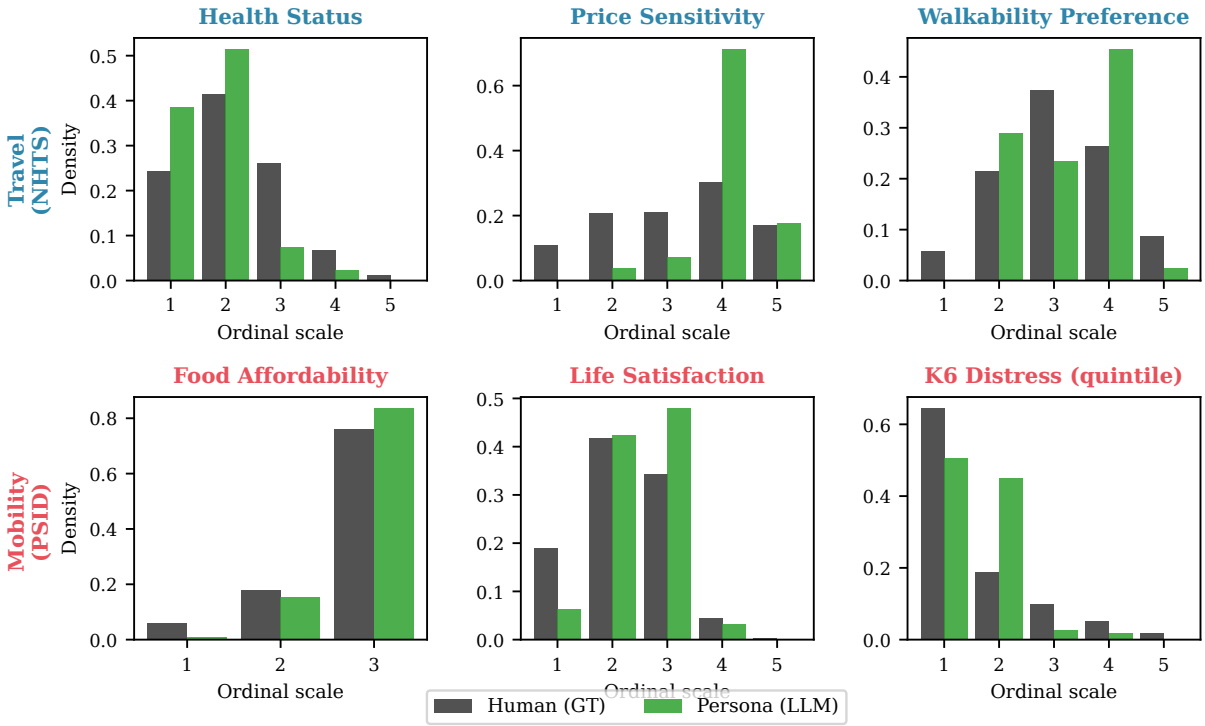

Figure 5: **Perception probe distributional fidelity.** Normalized histograms comparing human ground truth (dark) vs. synthesized persona responses (green) for all six holdout probes. Top row: Travel domain (NHTS). Bottom row: Mobility domain (PSID).

where $w_{ij} = (i - j)^2$ is the quadratic weight, O_{ij} is the observed confusion matrix, and E_{ij} is the expected matrix under independence.

Wasserstein Distance (EMD) To ensure the population of agents reproduces the aggregate public opinion distribution rather than collapsing to the mode, we calculate the Wasserstein Distance (Earth Mover's Distance) (Santurkar et al., 2023) between the cumulative distribution functions (CDF) of the human (U) and agent (V) responses:

$$W_1(U, V) = \sum_{k=1}^{K} |CDF_U(k) - CDF_V(k)| \quad (18)$$

where K is the number of ordinal Likert scale points. A lower Wasserstein distance indicates that the shape of the simulated opinion distribution closely matches the human ground truth.

Structural Correlation Alignment To verify that the model captures latent sociological mechanisms, we compute the *Structural Correlation Alignment* (Aher et al., 2023). For each domain, we form a correlation vector from all Spearman correlations between demographics and held-out probe responses. For NHTS, demographics are age, household income, tract population density, and vehicle count; for PSID, they are reference-person age, family income, home-ownership status, and self-rated health of the household head (RP_HEALTH). Each domain therefore contributes twelve demographic–probe pairs (4×3). Urbanicity was unavailable in our 2023 PSID extract; we include health because it appears in persona facts but is excluded from the held-out wellbeing probes. Let ρ_{obs} denote the human vector and ρ_{sim} the persona vector. The alignment score is:

$$\text{Alignment} = \text{Spearman}(\rho_{obs}, \rho_{sim}) \quad (19)$$

A high positive value indicates that the agents replicate the underlying relationships (e.g., higher income associating with lower food-affordability stress) found in the real world.

H Moderator Human Calibration Details

To ensure that the LLM judge reliably enforces constraints during intra-conversation control, we conducted a human calibration study. We recruited four graduate student annotators from different countries and with diverse gender backgrounds, all of whom had expertise in human decision-making. The annotators participated voluntarily and received no monetary compensation.

H.1 Annotation Rubric and Protocol

A panel of four human annotators evaluated the model outputs double-blind on a 1–3 scale (1=Poor, 2=Adequate, 3=Good) across four dimensions:

1. **Persona Consistency (per-member):** Evaluates whether the agent stays in character and reflects their socioeconomic profile without hallucinating false identities.

2. **Constraint Adherence (per-member):** Evaluates if the agents perfectly obey the physical, financial, and temporal limits of the household.

3. **Interaction Validity (holistic):** Evaluates the quality, flow, and realism of the multi-agent negotiation (e.g., dynamic deliberation vs. repetitive loops).

4. **Decision Quality (holistic):** Evaluates the logical soundness and mathematical accuracy of the final agreed-upon outcome.

Annotations were conducted in batches with a 20% cross-annotator overlap to monitor drift. Inter-Rater Reliability (IRR) was computed continuously to ensure alignment.

H.2 Inter-Rater Reliability (IRR) and Alignment Results

Table 10 reports the multi-rater agreement (Fleiss' κ (Fleiss, 1971)) and the average pairwise Quadratic Weighted Kappa (QWK (Cohen, 1968)) for the human annotators. Additionally, to demonstrate the successful transfer of these human preferences to the LLM judge, we compare the LLM–Human QWK before and after role-conditioned SFT on the Travel (Puget Sound) domain.

Dimension	Travel (Puget Sound)			Mobility (PSID)		
	Human QWK	**LLM (Before) QWK**	**LLM (After) QWK**	**Human QWK**	**LLM (Before) QWK**	**LLM (After) QWK**
Persona Consistency	0.446	0.133	0.450	0.847	0.139	0.490
Constraint Adherence	0.626	0.188	0.533	0.517	0.162	0.471
Interaction Validity	0.491	0.272	0.527	0.500	0.236	0.538
Decision Quality	0.845	0.506	0.749	0.729	0.460	0.648
Overall Average	**0.698**	**0.275**	**0.567**	**0.851**	**0.249**	**0.536**

Note: Overall human Fleiss' $\kappa = 0.620$ *(Travel) and* 0.853 *(Mobility).*

Table 10: **Moderator calibration inter-rater reliability.** Average pairwise QWK by rubric dimension. Travel (Puget Sound) includes LLM–Human QWK before and after role-conditioned SFT; Mobility (PSID) reports human agreement only.

Variable	Description
HHVEHCNT*	Number of vehicles in household
HHFAMINC*	Household income category
HOMEOWN*	Home ownership status
URBRUR	Urban vs rural household location
RAIL	Rail availability in household area
TRAVDAY	Travel day type (weekday vs weekend)
CHILDREN	Presence of children in household (under 18)
DRVRCNT	Number of licensed drivers in household
GASPRICE	Gas price on the travel day (cents)
GT1JBLWK	Indicator for having more than one job
HHSIZE*	Household size (number of members)
HH_HISP	Hispanic status of household respondent
HTEEMPDN	Employment density cluster (workers per square mile)
HTHTNRNT	Renter-occupied housing percentage cluster
HTPPOPDN	Population density cluster (people per square mile)
HTRESDN	Housing unit density cluster (units per square mile)
LPACT	Frequency of light/moderate physical activity (past week)
MSASIZE	Metropolitan Statistical Area (MSA) population size category
PC	Frequency of desktop/laptop internet use
PHYACT	Physical activity level (ordinal)
SPHONE	Frequency of smartphone internet use
YOUNGCHILD*	Presence of young child in household (typically age 0–5)
MSACAT_01	MSA $\geq$ 1M population, with rail
MSACAT_02	MSA $\geq$ 1M population, without rail
MSACAT_03	MSA $<$ 1M population
MSACAT_04	Not in an MSA
LIF_CYC_02	Household life cycle: 2+ adults, no children
LIF_CYC_03	Household life cycle: one adult, youngest child age 0–5
LIF_CYC_04	Household life cycle: 2+ adults, youngest child age 0–5
LIF_CYC_05	Household life cycle: one adult, youngest child age 6–15
LIF_CYC_06	Household life cycle: 2+ adults, youngest child age 6–15
LIF_CYC_07	Household life cycle: one adult, youngest child age 16–21
LIF_CYC_08	Household life cycle: 2+ adults, youngest child age 16–21
LIF_CYC_09	Household life cycle: one adult, retired, no children
LIF_CYC_10	Household life cycle: 2+ adults, retired, no children
AGE_U18_PCT*	Percent of household members under age 18
AGE_18_35_PCT*	Percent of household members age 18–35
AGE_35_54_PCT*	Percent of household members age 35–54
AGE_55_64_PCT*	Percent of household members age 55–64
AGE_65P_PCT*	Percent of household members age 65 and older
MALE_PCT*	Percent of household members who are male
WHITE_PCT*	Percent of household members who are White
MEDCOND_PCT	Percent of household members with medical conditions
WORKER_PCT*	Percent of household members who are employed
DRIVER_PCT*	Percent of household members who are licensed drivers
OCCAT_01_SalesService_PCT	Percent of household members in Sales/Service occupations
OCCAT_02_ClericalAdmin_PCT	Percent of household members in Clerical/Administrative occupations
OCCAT_03_ManualLabor_PCT	Percent of household members in Manual/Farming/Labor occupations
OCCAT_04_Professional_PCT	Percent of household members in Professional/Technical occupations
OCCAT_97_Other_PCT	Percent of household members in Other occupation categories
EDUC_OTHER_PCT*	Percent of household members with education below bachelor (or other non-degree category)
EDUC_BACHELOR_PCT*	Percent of household members with a bachelor degree
EDUC_GRAD_PCT*	Percent of household members with a graduate degree
STUDENT_PCT*	Percent of household members who are students
GT1JBLWK_PCT*	Percent of household members who have more than one job
POOR_HEALTH_PCT	Percent of household members reporting poor health
LPACT_MEAN	Mean light/moderate activity frequency across household members
PHYACT_LIGHT_PCT	Percent of household members with light activity level
PHYACT_MODERATE_PCT	Percent of household members with moderate activity level
PHYACT_VIGOROUS_PCT	Percent of household members with vigorous activity level
RIDESHARE_MEAN	Mean rideshare usage count across household members
PTUSED_MEAN	Mean public transit usage count across household members
BORNINUS_PCT	Percent of household members born in the United States

Table 11: Variable names and descriptions for NHTS 2017. Variables marked with "*" appear in both NHTS and Puget Sound.

Variable	Description
HHSIZE	Household size
NUM_CHILDREN_FU	Number of children in the family unit
AGE_YOUNGEST_CHILD	Age of the youngest child in the household
MARITAL_STATUS*	Marital status of the reference person
HOMEOWN*	Home ownership or housing tenure status
STATE_FIPS*	State location code
RACE_RP*	Race of the reference person
ETHNIC_RP*	Ethnic identification of the reference person
IMMIGRANT_SAMPLE*	Indicator for PSID immigrant sample membership
RP_AGE	Age of the reference person
SP_AGE	Age of spouse or partner
RP_SEX*	Sex of the reference person
SP_SEX*	Sex of spouse or partner
RP_HEALTH	Self-reported health of the reference person
SP_HEALTH	Self-reported health of spouse or partner
RP_EDUC_YEARS	Years of education of the reference person
SP_EDUC_YEARS	Years of education of spouse or partner
HHVEHCNT	Number of vehicles in household
HHFAMINC	Total household/family income
NEEDS_STANDARD_2022	Census needs standard for the household in 2022
INCOME_TO_NEEDS	Household income-to-needs ratio
CREDIT_CARD_DEBT	Household credit card debt
WEALTH_NO_EQUITY	Household wealth excluding home equity
BIRTH_RP_SP_2021	Birth-related measure for reference person/spouse in 2021
BIRTH_2021_FLAG*	Indicator for a birth in 2021
RP_WEEKS_LAIDOFF_2022	Weeks reference person was laid off in 2022
SP_WEEKS_LAIDOFF_2022	Weeks spouse/partner was laid off in 2022
RP_WEEKS_UNEMP_2022	Weeks reference person was unemployed in 2022
RP_WEEKS_UNEMP_2021	Weeks reference person was unemployed in 2021
SP_WEEKS_UNEMP_2021	Weeks spouse/partner was unemployed in 2021
RP_LUNG_LIMIT_ACTIVITY	Reference person activity limitation due to lung disease
RP_ARTHRITIS_LIMIT_ACTIVITY	Reference person activity limitation due to arthritis
SP_ARTHRITIS_LIMIT_ACTIVITY	Spouse/partner activity limitation due to arthritis
ANY_LAYOFF_2022*	Indicator for any household layoff in 2022
ANY_UNEMP_2022*	Indicator for any household unemployment in 2022
ANY_UNEMP_2021*	Indicator for any household unemployment in 2021
ANY_DISABILITY_LIMIT*	Indicator for any household disability/activity limitation
MAX_YEARS_EDUC	Maximum years of education in the household
MAX_NUM_JOBS_PY	Maximum number of jobs held in the prior year
PERSON_COUNT	Number of persons represented in household records
CHILDREN_PCT	Share of household members who are children
MALE_PCT	Share of household members who are male
SENIOR_PCT	Share of household members who are seniors
AGE_U18_PCT	Share of household members under age 18
AGE_18_35_PCT	Share of household members aged 18–35
AGE_35_54_PCT	Share of household members aged 35–54
AGE_55_64_PCT	Share of household members aged 55–64
AGE_65P_PCT	Share of household members aged 65 or older
EMPLOYED_PCT	Share of household members employed
RETIRED_PCT	Share of household members retired
STUDENT_PCT	Share of household members who are students
DISABLED_PCT	Share of household members with disability status
HEALTHY_PCT	Share of household members reporting good health
MULTI_JOB_PCT	Share of household members holding multiple jobs
EDUC_BACHELOR_PCT	Share of household members with a bachelor's degree
EDUC_GRAD_PCT	Share of household members with graduate education
COLLEGE_DEGREE_PCT	Share of household members with a college degree

Table 12: Processed PSID variables used for residential mobility prediction. Variables marked with * are treated as categorical variables in the preprocessing pipeline.