WHEN DECISION THEORIES EMPOWER AI AGENTS: LEARNING AND ORCHESTRATING THEORY–AGENT ALIGNMENT FOR PERSONALIZED DECISION-MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) hold great promise as personalized decision companions. However, using LLM agents for decision modeling also introduces new challenges: while LLMs offer richer and more flexible modeling capabilities, they often lack theoretical grounding, exhibit unstable reasoning behavior, and provide limited interpretability. This raises two central questions: how can AI agents be aligned with decision theories to support trustworthy reasoning, and how can the most appropriate theory-guided agents be selected or orchestrated for different tasks and personas? We address these questions through the lens of the *wisdom of crowds*: rather than searching for a single universal decision theory, we treat diverse decision theories as complementary behavioral experts. Building on this idea, we propose *Learning to Agentic Orchestration* (L2AO), a framework that orchestrates theory-guided AI agents for trustworthy and personalized decision modeling. L2AO consists of two key components: *Theory-guided Agent Learning* (TAL), which constructs and trains decision agents under different behavioral theories, and *Consensus-First Expert Adaptation* (CFEA), which coordinates learned agents through a democratic AI mechanism that prioritizes cross-theory consensus while adapting to task- and persona-specific differences. Experiments on five real-world survey-based decision datasets show that L2AO achieves state-of-the-art performance compared with utility-based, machine-learning, and LLM-based baselines, improving average accuracy by 15.0% over the strongest baseline. Beyond prediction, L2AO further reveals potential systematic biases that can arise when a single decision theory is embedded into an LLM agent to model diverse human decisions. The source code will be fully released upon paper acceptance.

1 INTRODUCTION

The same traveler may pay more to save ten minutes on a routine trip, yet avoid a faster route when an uncertain delay could mean missing a flight. Traveling with family may shift the choice once again. Such choices admit different behavioral explanations: utility trade-offs, sensitivity to potential losses, social context, or heterogeneous preferences (15; 13). Decades of decision research have translated these explanations into interpretable mathematical theories (10; 2). Yet having a rich collection of theories leaves a consequential question unresolved: which theory should guide the modeling of a particular decision? A theory that works well for routine travel may be less suitable for a health decision; one that captures the preferences of a population may miss the priorities of a particular person (22; 20; 32). Personalized decision modeling therefore depends on matching behavioral assumptions to the circumstances in which people choose.

Large language model (LLM) agents offer a promising means of connecting decision theories with rich descriptions of decision scenarios. They can process textual context, integrate numerical and semantic information, and adapt their reasoning to diverse tasks (41; 16; 44). However, this flexibility does not by itself ensure that an agent applies the right behavioral theory in the right setting. An LLM may generate a plausible explanation without consistently following the behavioral assumptions of a formal decision model (1). Even when an agent is grounded in a particular theory, that theory may be poorly suited to the task or decision-maker it is intended to model. Addressing this problem requires

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

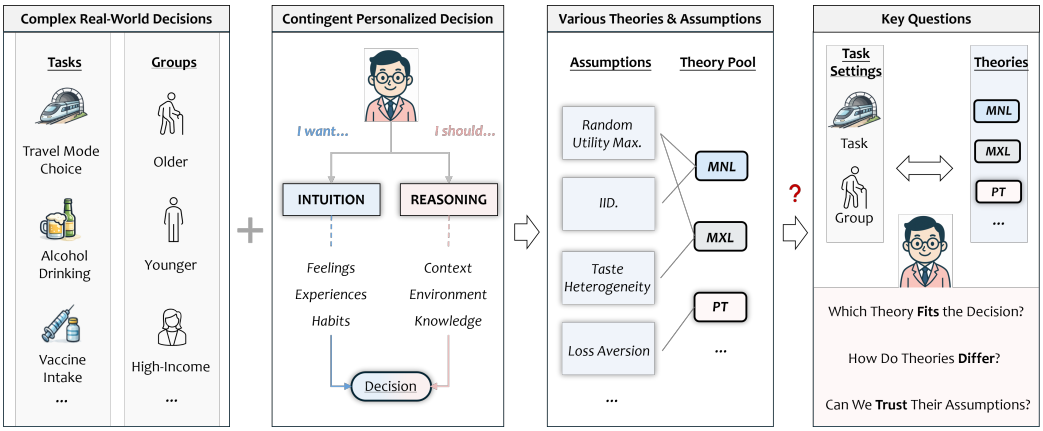

Figure 1: **Aligning Decision Theories with Decision Scenarios.** Human decisions vary across tasks, populations, and individuals, while decision theories encode different behavioral assumptions. This diversity raises a central challenge: how can different theories be matched to the decision scenarios they are best suited to model?

explicit connections between agent reasoning, measurable utilities, and observed choices, so that the suitability of different behavioral assumptions can be empirically evaluated (5).

We therefore investigate **how LLM agents can align different decision theories with the decision scenarios they are best suited to model**. We organize this problem into three research questions: *1) Theory grounding:* How can LLM agents reason under explicit decision-theoretic assumptions? *2) Scenario matching:* How can we identify the theories best suited to a given task and decision-maker profile? *3) Adaptive orchestration:* How can theory-guided agents be coordinated to apply suitable theories in each decision scenario?

Our approach treats decision theories as complementary behavioral perspectives. Utility-based models formalize trade-offs among attributes (28); prospect theory captures reference-dependent evaluations of gains and losses (21); models of dependence among alternatives represent relationships within a choice set (7); and heterogeneity-aware models accommodate variation in preferences (29). Instantiating these perspectives as theory-guided agents allows their predictions to be compared within a shared decision setting. Agreement indicates that a prediction is less sensitive to the choice among the represented theories, while disagreement identifies cases in which the underlying assumptions matter. These signals provide a basis for deciding when to rely on a shared prediction and when to invoke a more specialized behavioral perspective.

To operationalize this approach, we propose **L2AO**, a framework for learning and orchestrating LLM agents with decision theories for personalized human decision modeling. L2AO connects theory grounding with scenario matching and adaptive inference through two components:

- **Theory-guided Agent Learning (TAL).** TAL uses decision theories as explicit behavioral assumptions that guide agent reasoning and learning. Each agent combines symbolic utility modeling with LLM-based textual reasoning, linking a formal representation of preferences to the interpretation of contextual information. TAL jointly updates these two components using theory-specific numerical utility gradients and textual gradients derived from LLM feedback. This enables agents to learn under distinct theoretical assumptions while retaining a connection to measurable decision outcomes.
- **Consensus-First Expert Adaptation (CFEA).** CFEA coordinates the theory-guided agents at inference time. It uses their shared prediction when they agree and group–theory alignment to guide inference when they disagree. In this way, differences among agents inform the application of behavioral theories to particular decision scenarios.

We evaluate L2AO on five real-world survey-based decision-making tasks, where it improves average accuracy by 15.0% and weighted F1 by 12.2% over the strongest baseline. Specifically, L2AO

achieves weighted-F1 gains of 29.0% over the strongest machine-learning baseline, 33.5% over the best classical utility model, and 11.1% over the best single theory-guided agent. Our analysis further shows that L2AO not only improves prediction, but also exposes potential systematic differences in how behavioral theories represent population groups. For example, middle-income groups exhibit the highest theory disagreement, while high-income groups are more consensus-stable. These patterns suggest potential representational biases when a single decision theory is embedded into AI agents.

2 Related Work

Utility-based Choice Modeling. Modeling human decision-making has a long tradition in economics and behavioral science, where different discrete choice theories are built upon distinct *behavioral assumptions* about how individuals form preferences and evaluate alternatives. Classical models such as Multinomial Logit (MNL) assume stable utility maximization (28), while extensions such as Mixed Logit (MXL) introduce latent preference heterogeneity across individuals (29). Other formulations capture alternative mechanisms, including pairwise comparison and substitution effects (PCL) (7), and reference-dependent, nonlinear valuation under risk (PT) (21). Rather than forming a single unified account, these theories reflect different behavioral regimes, a phenomenon also observed in large-scale empirical studies of human choice (34). These behavioral differences arise not only across tasks but also across populations. Empirical studies in transportation and health economics have consistently shown that preference structures vary across demographic groups, motivating models that explicitly account for heterogeneity (11; 9). This means that a theory may perform well for one group but fail for another, not due to model inadequacy, but because its underlying assumptions are misaligned with the population it is applied to.

LLM-based Decision-making Modeling. Recent work has explored the integration of LLMs with utility-based decision models, aiming to combine structured statistical reasoning with rich contextual understanding. Prior studies have shown that LLMs can capture semantic and contextual factors beyond traditional discrete choice models (18; 26), while optimization frameworks such as TextGrad (46) enable iterative refinement of prompts and symbolic structures through textual feedback. Building on this paradigm, frameworks such as ATHENA (47) and DualMind (45) further integrate LLMs with symbolic utility representations to enable interpretable and flexible decision modeling. These approaches are conceptually aligned with cognitive theories such as dual-process theory (31), where symbolic utility representations capture stable, rational patterns, while LLMs provide complementary semantic reasoning. However, existing frameworks still relied on pre-defined theoretical assumptions, failing to capture heterogeneous decision mechanisms among tasks and groups.

Democratic AI for Decision-making. Democratic AI aggregates multiple agents or models to enhance decision reliability. Prior research on AI-mediated deliberation demonstrates that structured aggregation can extract consistent signals from conflicting inputs (24; 40). Notably, collective choice mechanisms such as generative voting have proven effective in mitigating stochastic variability and biases in individual LLM outputs (27). However, these approaches typically assume that disagreement arises from noise or uncertainty, and thus can be resolved through aggregation. This assumption breaks down when disagreement reflects fundamentally different *decision mechanisms*, rather than random variation (37; 14). In such cases, simple aggregation may obscure meaningful differences between underlying behavioral assumptions.

3 Method

We present **L2AO**, a theory-guided agent learning and orchestration framework for personalized decision modeling under heterogeneous behavioral assumptions. Rather than treating classical decision theories as standalone predictors, L2AO uses them as *assumption conditions* for training and organizing decision agents. The framework consists of two components: **Theory-guided Agent Learning (TAL) and Consensus-First Expert Adaptation (CFEA)**. In TAL, each theory defines an assumption-conditioned agent that jointly learns a symbolic choice model and a textual reasoning component. CFEA then orchestrates these trained agents at inference time by using consensus when agents agree and group–theory alignment when they disagree.

3.1 THEORY-GUIDED AGENT FOR PERSONALIZED DECISION-MAKING

Symbolic choice modeling. L2AO first formulates behavioral assumptions as explicit conditions that shape both utility-based choice modeling and LLM-based reasoning. Let x denote an individual decision context and $y \in \mathcal{Y}$ denote the observed choice. Let $a \in \mathcal{A}$ denote a behavioral assumption condition, such as MNL, MXL, PT, or PCL. For clarity, we describe the learning process within a fixed demographic group and omit the group index unless group-level orchestration is discussed. Under assumption condition a, L2AO first evaluates a symbolic utility representation $V(x;\theta)$, where θ denotes the utility parameters. The utility is then transformed into an assumption-conditioned choice distribution,

$$\pi_{x,a} = \Psi(V(x;\theta) \mid a)\,, \tag{1}$$

where $\pi_{x,a}$ denotes the probability distribution over candidate choices induced by assumption a, and $\Psi(\cdot \mid a)$ is the corresponding choice operator. Different assumptions define different ways of converting utility into choice probabilities. For example, MNL uses a logit-type utility comparison, MXL introduces preference heterogeneity, PT applies reference-dependent valuation, and PCL captures pairwise substitution patterns. The assumption-conditioned choice distribution is then used as a behavioral prior for LLM reasoning. Assumption mapping details are provided in Appendix A.1.

Textual reasoning. Following DualMind (45), L2AO adopts a dual-system reasoning structure that combines a symbolic system and a textual system. The symbolic system represents the utility function and the assumption-conditioned choice distribution $\pi_{x,a}$, whereas the textual system represents domain-specific knowledge K and describes the behavioral assumption in natural language. Both systems are integrated into a prompt template together with the textualized decision context $T(x)$. The LLM-based reasoning agent then generates a predicted choice and a confidence score conditioned on the behavioral assumption. This process is written as

$$(\hat{y}_{x,a}, c_{x,a}) = F_{\Theta}\left(T(x), K, \pi_{x,a} \mid a\right), \tag{2}$$

where $\hat{y}_{x,a}$ and $c_{x,a}$ denote the predicted choice and confidence score for case x under assumption condition a, Θ denotes the fixed LLM-based reasoning interface.

3.2 THEORY-GUIDED AGENT LEARNING

L2AO learns each assumption-conditioned decision path through two coupled updates: a *numerical utility-parameter update* and a *theory-aware symbolic&textual update*. The numerical update refines the parameters of the current utility function, while the symbolic&textual update revises the utility form and the knowledge set. These two updates iteratively improve the alignment of LLM agents with real-world human decisions under given theoretical assumption a.

Numerical utility-parameter update. Given the behavioral prior induced by assumption a, the numerical loss is defined as

$$\mathcal{L}^{\mathrm{num}}(a) = -\log \pi_{x,a}(y). \tag{3}$$

In the iteration step t, the utility parameters are updated with step size η:

$$\theta^{t+1} = \theta^{t} - \eta \nabla_{\theta} \mathcal{L}^{\mathrm{num}}(a). \tag{4}$$

Theory-aware symbolic&textual update. In parallel, L2AO constructs theory-aware textual gradients to revise the utility-function form and the knowledge set. The textual gradient is natural-language feedback that specifies how the textual components of an LLM system should be updated to improve its outputs (46). In L2AO, this feedback is theory-aware in two ways: 1) it is conditioned on the behavioral assumption, and 2) it incorporates numerical theory-aware signals from the symbolic choice model. The textual gradient under assumption a can be represented as:

$$G(a) = \Lambda\left(x, y, \hat{y}, \mathcal{L}^{\mathrm{num}}, \nabla_{\theta}\mathcal{L}^{\mathrm{num}}, V^{t}, K^{t} \mid a\right), \tag{5}$$

where $\Lambda(\cdot \mid a)$ denotes an LLM-based textual feedback generator. It converts both language-level prediction errors and numerical gradient signals into unified and structured revision instructions. These instructions may revise the utility form, add or remove behavioral factors, adjust how variables

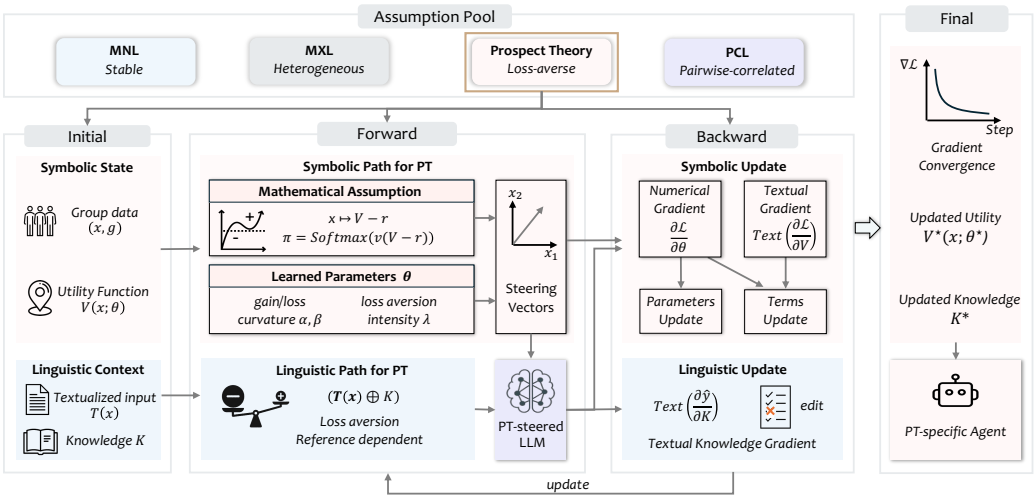

Figure 2: **Theory-Guided Agent Learning in L2AO.** We take PT as an example. During the forward process, each behavioral assumption defines both a symbolic choice path and a linguistic reasoning path. These two paths are integrated in a theory-steered LLM predictor. During the backward process, numerical utility gradients from the theory-specific symbolic path and textual knowledge gradients from the linguistic path are integrated within a unified textual update framework, jointly updating the symbolic state and linguistic context. After optimization, each behavioral assumption induces a trained theory-conditioned predictor.

enter the utility expression, or update the textual knowledge set. The utility form and knowledge set are then updated by an LLM optimizer $\mathcal{T}$:

$$\left(V^{t+1}, K^{t+1}\right) = \mathcal{T}\left(V^t, K^t, G(a) \mid a\right). \tag{6}$$

The numerical gradient updates *how much* each factor matters by refining the utility parameters. The theory-aware textual gradient updates *what form* the utility and knowledge should take by revising the symbolic structure and behavioral reasoning rules. Because both updates are conditioned on a, each reasoning path remains consistent with its theoretical foundation.

Theory-guided agents. After training, each assumption condition defines an optimized theory-specific agent. For a new case x, the corresponding theory-specific agent is then written as

$$E^\star(x \mid a) = F_\Theta\left(T(x), K^\star, \pi^\star_{x,a} \mid a\right). \tag{7}$$

where $\pi^\star_{x,a} = \Psi(V^\star(x; \theta^\star) \mid a)$ is the optimized assumption-conditioned choice distribution.

3.3 Consensus-First Expert Adaptation Mechanism

L2AO orchestrates the optimized theory-conditioned agents through a consensus-first democratic mechanism. Different behavioral assumptions provide complementary behavioral views of the same decision context. Agreement among assumptions indicates a stable cross-theory signal, whereas disagreement suggests that the prediction is sensitive to the underlying behavioral assumption. In the latter case, L2AO relies on the assumption that is most aligned with the demographic group.

Consensus-first democratic decision. For a test case x, each optimized theory-conditioned agent $E^\star(x \mid a)$ produces a predicted choice $\hat{y}_{x,a}$ and a confidence score $c_{x,a}$ for $a \in \mathcal{A}$. L2AO defines the consensus prediction as the most frequently predicted choice across assumptions: $\hat{y}_{\text{cons}}(x) = \text{mode}_{a\in\mathcal{A}}\, \hat{y}_{x,a}$, where the mode is taken over the set of theory-conditioned predictions $\{\hat{y}_{x,a} : a \in \mathcal{A}\}$. The strength of this consensus is measured by confidence-weighted agreement:

$$\rho(x) = \frac{\sum_{a\in\mathcal{A}} c_{x,a} \mathbb{I}_{\hat{y}_{x,a}=\hat{y}_{\text{cons}}(x)}}{\sum_{a\in\mathcal{A}} c_{x,a}}. \tag{8}$$

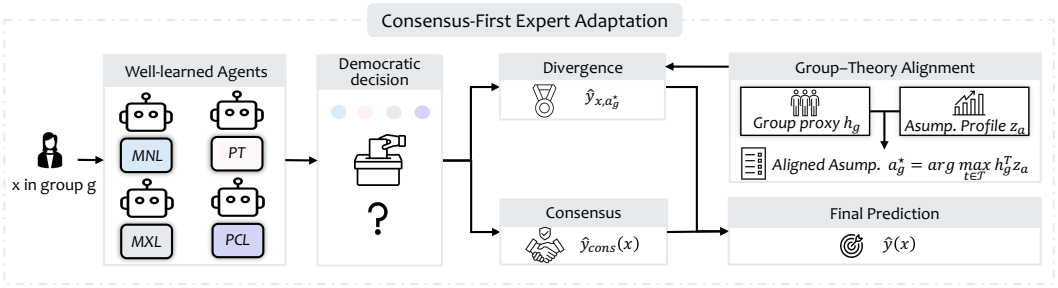

Figure 3: **Consensus-First Expert Adaptation in L2AO.** At inference time, multiple theory-conditioned agents first generate predictions in parallel. When their outputs agree, the framework adopts the consensus prediction; otherwise, it defers to group–theory alignment and selects the assumption whose profile best matches the group proxy.

The score $\rho(x) \in [0, 1]$ measures how much confidence mass supports the consensus prediction. A high value indicates that the decision is robust across theoretical assumptions, whereas a low value indicates cross-theory disagreement. If $\rho(x) \geq \tau$, L2AO adopts the consensus prediction. If $\rho(x) < \tau$, the consensus is considered weak, and L2AO selects the theoretical assumption most aligned with the demographic group. Details of τ and consensus mechanism are provided in Appendix A.7.

Group-theory alignment. This group-theory alignment is computed as

$$a_g^\star = \arg\max_{a \in \mathcal{A}} \left(h_g^T z_a \right), \tag{9}$$

where h_g is the behavioral proxy of demographic group g, and z_a is the learned profile of assumption a. The proxy h_g summarizes the validation behavior of group g, while z_a summarizes the behavioral pattern captured by assumption a. Details of h_g, z_a, and their normalization are provided in Appendix A.8. Thus, $a_g^\star$ identifies the assumption whose learned behavioral structure is most compatible with the group.

Final prediction. The final prediction is selected between consensus and group-aligned prediction.

$$\hat{y}(x) = \underbrace{\mathbb{I}_{\rho(x) \geq \tau}\, \hat{y}_{\text{cons}}(x)}_{\text{consensus}} + \underbrace{\mathbb{I}_{\rho(x) < \tau}\, \hat{y}_{x,a_g^\star}}_{\text{group-aligned agent}}. \tag{10}$$

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate the proposed framework on five real-world decision-making datasets from survey: **1)** Swiss Metro travel mode choice (**SWM**) (3; 35; 42), with alternatives *Train, Car, Swissmetro*; **2)** COVID-19 vaccine uptake (**VAC**) (25), with alternatives *Unvaccinated, Vaccinated, Booster*; **3)** job choice during COVID-19 (**JOB**) (30), with alternatives *Worked for salary, Self-employed, Did not work*; **4)** marijuana use recency (**MAJ**) and **5)** alcohol use recency (**ALC**) from NSDUH (39), with alternatives *past 30 days, past 12 months, >12 months*. To capture population heterogeneity, each dataset is partitioned into demographic groups defined by age (*young, middle-aged, older*) and income (*below, middle, above*). For the VAC dataset, the income-related variable contains *below-median*, *above-median*, and *non-response* categories. All theory-based models are estimated at the group level, and consensus-gated inference is performed at the same level of granularity.

Models and baselines. Our proposed framework uses four representative and behaviorally distinct assumptions, Multinomial Logit (MNL) (28), Paired Combinatorial Logit (PCL) (7), Prospect Theory (PT) (21), and Mixed Logit (MXL) (29), as its coordinated assumption set. We compare against three groups of baselines. *(i)* Standard machine-learning methods: MLP (36), XGBoost (4), and symbolic

Table 1: **Performance comparison across 5 datasets.** LLM-V refers to vanilla LLM (GPT-4o-mini). The **All** columns report the average F1 and Acc across five datasets.

	Method	ALC		JOB		MAJ		SWM		VAC		All	
		F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑
Machine Learning	GPlearn (38)	0.349	0.367	0.489	0.492	0.220	0.242	0.465	0.466	0.594	0.600	0.423	0.433
	PySR (8)	0.435	0.430	0.504	0.517	0.445	0.492	0.615	0.620	0.563	0.595	0.512	0.531
	XGBoost (4)	0.430	0.435	0.522	0.516	0.527	0.496	0.631	0.627	0.504	0.571	0.523	0.529
	MLP (36)	0.418	0.403	0.432	0.471	0.577	0.563	0.617	0.613	0.597	0.610	0.528	0.532
Utility-based	MNL (28)	0.379	0.364	0.431	0.471	0.437	0.391	0.500	0.530	0.586	0.585	0.467	0.468
	MXL (29)	0.369	0.382	0.415	0.471	0.480	0.434	0.344	0.462	0.549	0.566	0.431	0.463
	PT (21)	0.372	0.371	0.386	0.443	0.490	0.445	0.525	0.548	0.616	0.624	0.478	0.486
	PCL (7)	0.400	0.404	0.562	0.588	0.465	0.426	0.517	0.548	0.605	0.626	0.510	0.518
	NL (43)	0.387	0.371	0.432	0.443	0.416	0.371	0.505	0.538	0.582	0.580	0.464	0.461
	RRM (6)	0.409	0.392	0.446	0.475	0.403	0.359	0.487	0.516	0.584	0.585	0.466	0.465
LLM-based	LLM-V	0.296	0.343	0.256	0.246	0.083	0.195	0.489	0.548	0.351	0.463	0.295	0.359
	TextGrad (46)	0.293	0.307	0.418	0.389	0.063	0.137	0.521	0.552	0.537	0.527	0.366	0.382
	Athena (47)	0.224	0.278	0.399	0.460	0.522	0.535	0.434	0.498	0.593	0.605	0.434	0.475
	DualMind (45)	0.469	0.488	0.645	0.623	0.633	0.629	0.677	0.681	0.610	0.589	0.607	0.602
L2AO (ours)		**0.552**	**0.590**	**0.774**	**0.762**	**0.718**	**0.762**	**0.694**	**0.695**	**0.669**	**0.649**	**0.681**	**0.692**

regression methods including GPlearn (38) and PySR (8). *(ii)* Theory-based methods: each decision theory applied independently as a standalone predictor. *(iii)* LLM-based methods: vanilla LLM; TextGrad (46), which optimizes text-based components through textual feedback; Athena (47), which combines group-level symbolic utility discovery with individual-level semantic template adaptation; and DualMind (45), which coordinates an automated iterative refinement between utility functions and LLM-based reasoning.

Implementation details. For all LLM-based prediction and final inference, we use GPT-4o-mini as the forward model (19). GPT-4o is used as the backward model to produce textual feedback and update utility–knowledge representations. Final decisions for all optimized theory-conditioned agents are made from the GPT-4o-mini forward prediction. TAL uses a fixed 16-step outer optimization budget with validation-F1 checkpointing, rollback, and rejection of invalid symbolic updates. Full implementation details are provided in Appendix A.

Evaluation metrics. We evaluate all methods using weighted F1 and accuracy computed over the test set. Performance metrics for L2AO are reported as the mean across 3 independent runs.

4.2 Experiments and Results

L2AO substantially improves LLM-based human decision modeling across diverse real-world tasks. Table 1 compares L2AO with machine-learning, utility-based, and LLM-based baselines on five real-world decision-making datasets. L2AO achieves the state-of-the-art F1 and accuracy on all datasets, with an overall average F1 of 0.681 and accuracy of 0.692. It improves F1 over the strongest machine-learning baseline, MLP, by 29.0%, over the best classical utility model, PCL, by 33.5%. Compared with LLM-based decision-modeling baselines, L2AO improves average F1 by 130.8% over direct LLM prediction, 86.1% over TextGrad, 56.9% over Athena, and 12.2% over DualMind.

TAL and CFEA jointly improve adaptation to heterogeneous real-world decision scenarios. We compare L2AO with single-theory agents to separate the contribution of TAL from that of CFEA. Each single-theory agent instantiates the learning stage of L2AO without CFEA: it is trained under one behavioral assumption and used independently at inference time. As shown in the left panel of Figure 4, two observations emerge. First, each theory-guided agent performs strongly on some tasks but not others: PCL is competitive on ALC, MXL performs well on JOB and MAJ, MNL is strong on SWM, and PT remains competitive on VAC. Second, CFEA enables L2AO to consistently outperform all single-theory agents, improving average F1 across datasets by 11.1%–15.4%. Together, these results clearly separate two sources of gain: TAL turns each assumption into a task-specialized

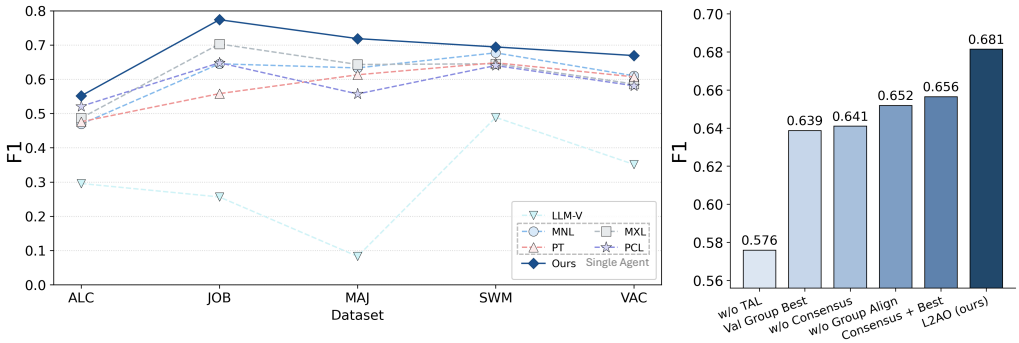

Figure 4: **Theory Orchestration Improves Decision Modeling Beyond Single-Theory Agents. Left**: Dataset-level F1 scores of LLM-Vanilla, four single-theory agents guided by MNL, MXL, PT, and PCL, and L2AO. **Right**: Ablation study of L2AO. *w/o TAL* removes the LLM-driven symbolic and textual updates in TAL while retaining numerical theory fitting and the same CFEA inference. *Val group best* directly applies the validation-selected best theory per group; *w/o Consensus* removes consensus-first democratic decision and directly uses the group-aligned theory-conditioned agent; *w/o Group Align* removes the group–theory alignment and randomly selects a theory-conditioned agent when agents disagree; *Consensus + Best* replaces group–theory alignment with the validation-selected best theory for each group when agents disagree.

decision agent, while CFEA coordinates these agents, leveraging the complementary strengths of different theories.

4.3 Ablation Study

The right panel of Figure 4 shows the ablation study of L2AO, with all results averaged across the five datasets. **Removing the LLM-driven symbolic and textual updates in TAL while retaining numerical theory fitting and the same CFEA inference reduces the average F1 from 0.681 to 0.576 (↓15.4%), demonstrating that TAL contributes substantially beyond numerical optimization alone.** Removing the consensus-first democratic decision step (*w/o Consensus*) causes a 5.9% F1 drop, yielding results comparable to directly applying the validation-selected best theory per group (*Val Group Best*). This indicates that group-level selection alone provides limited gains over static strategies. Using only consensus-first democratic decision and randomly selecting a theory-conditioned agent when agents disagree (*w/o Group Align*) yields stronger performance (↓4.3% F1), showing that cross-theory agreement provides a stable shared decision signal. However, it still underperforms the full model because it cannot resolve high-disagreement cases where theories exhibit conflicting valuation biases. We further compare with *Consensus + Best*, which replaces group-theory alignment with simple validation-based theory selection. While this improves over single-stage variants, it remains 3.7% below the full model, indicating that naive group-level selection cannot fully capture heterogeneous decision patterns. The gain of the full model over this variant suggests that group alignment is particularly effective in high-disagreement regions.

We further examine the robustness of L2AO to theory-pool composition, demographic-group construction, and independently regenerated data splits in Appendix E, where the main performance advantage remains consistent across these alternative settings.

4.4 Group–Theory Alignment Analysis

L2AO reveals theory-task and theory-group alignment patterns. The theory distribution in the left panel of Figure 5 shows clear and interpretable alignments. MNL is more often selected in SWM, covering 4/9 demographic groups. This suggests that transportation choices are well captured by stable utility trade-offs. MXL appears more often in MAJ, covering 6/9 demographic groups, where group-level decision patterns are more heterogeneous. PCL is more visible in ALC, covering 6/9 groups, where adjacent use-recency categories are behaviorally closer. Notably, PT is more frequently selected for older groups, covering 4/7 PT-aligned demographic groups, indicating greater empirical

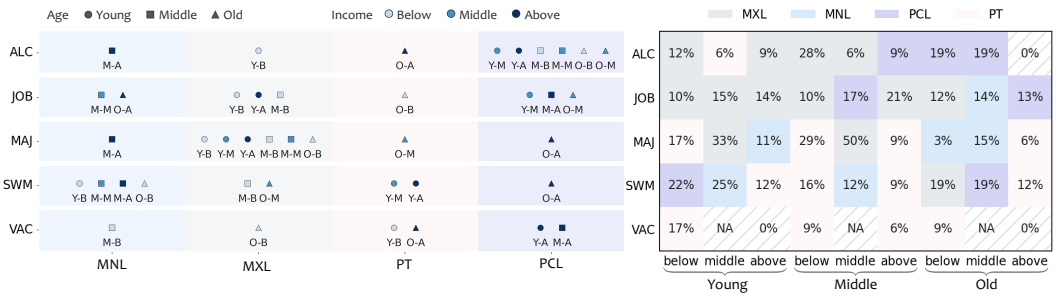

Figure 5: **Analysis of single-theory agents and group-level alignment. Left**: Each marker denotes the best-performing single-theory agents for an age–income subgroup, where labels indicate demographic groups (Y/M/O: young/middle/old; B/M/A: below/middle/above income). **Right**: Theory selection patterns on non-consensus cases. Rows denote datasets and columns denote demographic groups. Each cell reports the proportion of test cases that enter the non-consensus regime, with color indicating the selected aligned assumption. NA indicates unavailable groups in VAC.

compatibility between these groups and the PT-conditioned decision structure. These results show that L2AO not only improves prediction, but also helps interpret which behavioral assumptions fit different tasks and groups.

L2AO exposes systematic representational differences of behavioral theories when decision theories are adapted into AI agents. As shown in the right panel of Figure 5, middle-income groups have the highest pooled non-consensus rate across available task–group cells, reaching 20.2%, compared with 13.5% for low-income groups and 8.7% for high-income groups. These patterns indicate that decisions of middle-income groups are less often resolved by consensus and require more adaptive group–theory alignment, whereas high-income groups are generally more consensus-stable and can be captured more consistently by multiple theory-guided agents. At the task level, the strongest disagreements occur in MAJ, especially for young and middle-aged middle-income groups, with non-consensus rates of 33% and 50%, respectively, and in SWM for young middle-income groups, with a non-consensus rate of 25%. Together, these patterns suggest that embedding a single decision theory into an AI agent may introduce both group and task-level representational biases.

Take away. Older groups are more frequently matched with PT-conditioned agents, while middle-income groups are the most theory-contested and high-income groups are the most consensus-stable.

5 Conclusion

We present *Learning to Agentic Orchestration* (L2AO), a framework for orchestrating theory-guided AI agents in heterogeneous human decision modeling. Within L2AO, Theory-guided Agent Learning (TAL) integrates behavioral assumptions into LLM-agent reasoning and learning, while Consensus-First Expert Adaptation (CFEA) orchestrates theory-guided agents through a democratic consensus mechanism. Across five real-world decision datasets, L2AO consistently outperforms classical utility models, machine-learning baselines, LLM-based baselines, and single-theory agents, improving the average accuracy by 15.0% over the strongest baselines. L2AO further reveals that relying on a single theory-guided agent can introduce systematic biases across groups and tasks. Overall, L2AO offers a theory-grounded path toward trustworthy AI agents that can adapt to heterogeneous human decisions while improving model stability, interpretability, and behavioral transparency.

Limitations. The current implementation of L2AO has two major limitations. First, L2AO incurs relatively high computational cost, as it requires training multiple theory-conditioned agents and running multiple agents in parallel during inference. Second, L2AO currently relies on a predefined theory pool, so its effectiveness may be influenced by the coverage and relevance of the selected behavioral theories.

AI USE STATEMENT

In this work, we used generative AI tools to assist with portions of the experimental code development, result visualization, literature search, and paper polishing. For AI-assisted experimental code and figure generation, we manually reviewed the code structure and inspected intermediate and final outputs at each stage to verify correctness and consistency with the intended experimental procedure. We also used generative AI tools to identify potentially relevant literature; all suggested references were manually verified against the original sources to confirm their existence, relevance, and accuracy before being cited. Generative AI tools were additionally used to improve the clarity, grammar, and presentation of the manuscript. All AI-assisted content and artifacts were reviewed by the authors. We take full responsibility for the final content of this work, including all text, claims, code, experimental results, figures, and references produced with the aid of generative AI.

ETHICS STATEMENT

L2AO is intended for modeling and analyzing heterogeneous human decision patterns. By comparing multiple decision theories and explicitly identifying consensus and disagreement, the framework can improve transparency in behavioral modeling across domains such as transportation, public health, socioeconomic decisions, and substance-use behavior. All datasets are secondary public-use or publicly released research datasets. We did not recruit participants or attempt re-identification, and we use the data only under their stated access conditions and for aggregate behavioral modeling.

The use of demographic attributes such as age and income requires careful interpretation. **Group–theory compatibility reflects observed group-level decision distributions and should not be interpreted as an individual psychological diagnosis, causal relationship, or basis for differential treatment.** Such associations should not be used for employment screening, healthcare or insurance decisions, targeted persuasion, resource allocation, or similar consequential applications. Reporting theory disagreement is intended to reveal representational heterogeneity that could otherwise remain hidden under a single assumed behavioral model. However, misuse of these patterns may still introduce demographic profiling, stereotyping, fairness, or privacy concerns (12; 23).

L2AO also relies on LLM-based reasoning and may inherit stochasticity, bias, or instability from the underlying models (33; 17). Consequential deployment would therefore require domain-specific validation, fairness and robustness auditing, privacy safeguards, and human oversight. The subgroup findings in this study should be treated as descriptive and context-specific rather than generalized beyond the studied populations and decision settings.

REFERENCES

[1] Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. Faithfulness vs. plausibility: On the (Un)Reliability of explanations from large language models, 2024.

[2] Sudeep Bhatia and Lisheng He. Machine-generated theories of human decision-making. *Science*, 372(6547):1150–1151, 2021.

[3] Michel Bierlaire, Kay Axhausen, and Georg Abay. The acceptance of modal innovation: The case of swissmetro. In *Swiss transport research conference*, 2001.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *KDD*, 2016.

[5] Xiusi Chen, Shanyong Wang, Cheng Qian, Hongru Wang, Peixuan Han, and Heng Ji. DecisionFlow: Advancing large language model as principled decision maker. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16668–16692. Association for Computational Linguistics, 2025.

[6] Caspar G Chorus, Theo A Arentze, and Harry JP Timmermans. A random regret-minimization model of travel choice. *Transportation Research Part B: Methodological*, 42(1):1–18, 2008.

[7] Chunghsin Chu. A paired combinatorial logit model for travel demand analysis. *Proceedings of the 5th World Conference on Transport Research*, 4:295–309, 1989.

[8] Miles Cranmer. Interpretable machine learning for science with pysr and symbolicregression.jl, 2023.

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

[10] Donald L. DeAngelis and Stephanie G. Diaz. Decision-making in agent-based modeling: A current review and future prospectus. *Frontiers in Ecology and Evolution*, Volume 6 - 2018, 2019.

[11] Dick Ettema, Tommy Gärling, Lars E. Olsson, Margareta Friman, and Sjef Moerdijk. The road to happiness: Measuring dutch car drivers' satisfaction with travel. *Transport Policy*, 27:171–178, 2013.

[12] Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1), 2024.

[13] Erin B Fitz, Dominik A Stecuła, Matthew P Hitt, and Kyle L Saunders. Objective numeracy exacerbates framing effects from decision-making under political risk. *Scientific Reports*, 14(1):10473, 2024.

[14] Cary Frydman and Salvatore Nunnari. Coordination with cognitive noise. 2021.

[15] Baocheng Geng, Qunwei Li, and Pramod K. Varshney. Human decision making with bounded rationality. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5493–5497, 2022.

[16] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges, 2024.

[17] Joschka Haltaufderheide and Robert Ranisch. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *npj Digital Medicine*, 7(1):183, jul 2024.

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

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[20] Martin Höppner. Models in decision-making under risk and uncertainty. *Journal of Economic Surveys*, 40(1):304–320, 2026.

[21] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

[22] Evanthia Kazagli and Matthieu de Lapparent. A discrete choice modeling framework of heterogenous decision rules accounting for non-trading behavior. *Journal of Choice Modelling*, 48:100413, 2023.

[23] Alina Köchling and Marius Claus Wehner. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, 13(3):795–848, nov 2020.

[24] Raphael Koster et al. Human-centred mechanism design with democratic ai. *Nature Human Behaviour*, 6:1396–1411, 2022.

[25] Jeffrey V. Lazarus, Katarzyna Wyka, Trenton M. White, Camila A. Picchio, Lawrence O. Gostin, Heidi J. Larson, Kenneth Rabin, Scott C. Ratzan, Adeeba Kamarulzaman, and Ayman El-Mohandes. A survey of covid-19 vaccine acceptance across 23 countries in 2022. *Nature Medicine*, 29(2):366–375, 2023.

[26] Tianming Liu, Manzi Li, and Yafeng Yin. Aligning llm with humans for travel choices: A persona-based embedding learning approach. *Transportation Science*, 60(5):806–827, 2026.

[27] Srijoni Majumdar, Edith Elkind, and Evangelos Pournaras. Generative ai voting: fair collective choice is resilient to llm biases and inconsistencies. *EPJ Data Science*, 15(1):24, 2026.

[28] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, 1973.

[29] Daniel McFadden and Kenneth Train. Mixed mnl models for discrete response. *Journal of Applied Econometrics*, 15(5):447–470, 2000.

[30] Laura Montenovo, Xuan Jiang, et al. Determinants of disparities in covid-19 job losses. *Demography*, 59(3):827–855, 2022.

[31] Allen Newell. *Unified Theories of Cognition*. Harvard University Press, 1990.

[32] Marino Pagan, Vincent D. Tang, Mikio C. Aoi, Jonathan W. Pillow, Valerio Mante, and Carlos D. Brody. Individual variability of neural computations underlying flexible decisions. *Nature*, 639(8052):421–429, mar 2025.

[33] Tiago P. Pagano, Rafael B. Loureiro, Fernanda V. N. Lisboa, Rodrigo M. Peixoto, Guilherme A. S. Guimarães, Gustavo O. R. Cruz, Maira M. Araujo, Lucas L. Santos, Marco A. S. Cruz, Ewerton L. S. Oliveira, Ingrid Winkler, and Erick G. S. Nascimento. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1), 2023.

[34] Joshua C Peterson, David D Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L Griffiths. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214, 2021.

[35] Huy Pham, Xuan Jiang, and Cong Zhang. Causality and advanced models in trip mode prediction: Interest in choosing swissmetro. 2022.

[36] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 1986.

[37] Benjamin Scheibehenne, Jörg Rieskamp, and Eric-Jan Wagenmakers. Testing adaptive toolbox models: A bayesian hierarchical approach. *Psychological Review*, 120(1):39–64, 2013.

[38] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 2009.

[39] Substance Abuse and Mental Health Services Administration. National survey on drug use and health (nsduh), 2023. https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health.

[40] Michael Henry Tessler, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin Chadwick, Raphael Koster, et al. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719):316–321, 2024.

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

[42] Xuli Wen and Xin Chen. A new breakthrough in travel behavior modeling using deep learning: A high-accuracy prediction method based on a cnn. *Sustainability*, 17(2):738, 2025.

[43] H. C. W. L. Williams. On the formation of travel demand models and economic evaluation measures of user benefit. *Environment and Planning A*, 1977.

[44] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Wei Xu, Shihan Dou, Junjie Ye, Guiyang Hou, Yan Liu, Bin Gui, Haoxiang Fen, Yongzu Cano, Zhengando Wen, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

[45] Hao Frank Yang, Yang Zhao, Yichen Lin, and Takeru Igusa. Thinking, fast and slow in ai agents for personalized decision-making. 2026.

[46] Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. Optimizing generative ai by backpropagating language model feedback. *Nature*, 639(8055):609–616, 2025.

[47] Yibo Zhao, Yang Zhao, Hongru Du, and Hao Frank Yang. Personalized decision modeling: Utility optimization or textualized-symbolic reasoning. *Advances in Neural Information Processing Systems*, 38:89317–89364, 2026.

A IMPLEMENTATION DETAILS

A.1 ASSUMPTION-SPECIFIC CHOICE MAPS

For each group g and assumption a, the utility expression produces alternative-level utilities

$$V_{g,a}(x) = \big(V_{g,a,1}(x), \ldots, V_{g,a,m}(x)\big).$$

Unlike a fixed utility-steering framework, L2AO uses a theory-specific probabilistic map

$$P_a(j \mid x) = \Psi_a\big(V_{g,a}(x); \theta_a\big),$$

where Ψ_a is determined by the behavioral assumption and θ_a denotes its assumption-specific parameters.

For the four assumptions used in our framework, the symbolic choice maps are summarized as follows:

$$\Psi_{\mathrm{MNL}}: \quad P(j \mid x) = \mathrm{softmax}_j(V),$$

which assumes stable utility maximization with fixed marginal effects. For mixed logit,

$$\Psi_{\mathrm{MXL}}: \quad P(j \mid x) \approx \frac{1}{R}\sum_{r=1}^{R} \mathrm{softmax}_j\big(V^{(r)}\big),$$

where random coefficients induce heterogeneous preference realizations. For prospect theory,

$$\Psi_{\mathrm{PT}}: \quad P(j \mid x) = \mathrm{softmax}_j(V^{\mathrm{PT}}), \qquad V_j^{\mathrm{PT}} = v(V_j - r),$$

where $v(\cdot)$ applies reference-dependent and loss-sensitive valuation. For paired combinatorial logit,

$$\Psi_{\mathrm{PCL}}: \quad P_l = \sum_{k \neq l} P_{l|lk} P_{lk},$$

where each pair of alternatives has its own nesting parameter and the final probability aggregates pairwise substitution structures.

A.2 THEORY-CONDITIONED GRADIENT AND PROMPT DETAILS

We summarize, for each theory used in experiments, (i) the parameter set, (ii) components returned by `grad_nll_one` (used when textual gradients are mixed with numerical sensitivity to the fitted model), (iii) differences among general prompt, and (iv) Prospect Theory (PT)–specific knowledge and utility constraints. All models share deterministic utilities $V_j(\mathbf{x}; \theta)$ per alternative j, written as symbolic expressions with learnable coefficients; only the map $V \mapsto P(\cdot \mid \mathbf{x})$ differs.

A.3 PARAMETER SETS

MNL. Let $\mathcal{J}$ be the alternative set. Parameters are every scalar coefficient appearing in the symbolic utilities after alternative-specific prefixing. We denote this set by θ.

MXL. In addition to θ, for each random taste feature f in `random_features` the implementation maintains a population mean μ_f and a log-scale $\log \sigma_f$ for the standard deviation. Simulated choice probabilities use n-draws Monte Carlo (or Halton) draws; parameters are $(\theta, \{\mu_f, \log \sigma_f\}_{f \in \mathcal{F}})$.

PCL. Utility parameters θ_{util} are as in MNL. For each unordered pair of alternatives $\{k, \ell\}$ there is a nesting parameter $\lambda_{k\ell} \in (0, 1]$ (clamped during the forward pass).

PT. Utility parameters θ_{util} come from the same MNL backbone. The value function applies $v(x) = x^\alpha$ for $x \geq 0$ and $v(x) = -\lambda(-x)^\beta$ for $x < 0$ to centered scores $V_j - r$, with softmax over $v(V_j - r)$. Learned global parameters are α, β, λ.

A.4 UTILITY OPTIMIZER PROMPTS

Table 2 contrasts the *system* prompts that instruct the evaluator LLM how to rewrite the utility block under TextGrad. Task-specific *constraints* (feature lists, JSON schema, term limits) are supplied separately.

Table 2: Utility optimization system prompts in our implementation.

Theory	Role (conceptual)
MNL, MXL	Generic "improve utility functions"; states that revision uses (1) feedback from other variables and (2) numerical gradients of the utility. Output: `num_utility` JSON objects (one formula per alternative) in tagged blocks.
PCL	Identifies the model as PCL: value functions V_j and pairwise nest parameters $\lambda_{k\ell}$; probability $P_\ell = \sum_{k \neq \ell} P_{\ell\|k\ell}\, P_{k\ell}$. Same output tagging as MNL; numerical gradients are PCL NLL.
PT	Identifies PT logit: user edits *raw* scores V_j; mapping $v(V_j - r)$ then softmax. Tells the optimizer to emphasize cross-alternative spread in V_j relative to the reference (gains vs. losses). Same tagged JSON output format.

A.5 Prompt Templates and Theory-Specific Constraints

We use structured prompts for both forward prediction and backward optimization. All theory-conditioned agents share the same high-level prompt interface, while the <UTILITY> block and the optimizer instructions are specialized according to the behavioral assumption. Below we show representative PT prompt templates. The templates are shortened for readability while preserving the system role, input blocks, theory-specific constraints, and output format.

Prompt A.5.1: PT Forward Prediction Template

```
[SYS] You are a {TASK_NAME} Decision Assistant.
Your task is to predict the individual's final choice.

You will receive the following structured blocks:

<KNOWLEDGE>
Task-specific behavioral knowledge and instructions on how to use it.
</KNOWLEDGE>

<PROFILE>
Observed individual features, demographic attributes, and contextual
    variables.
</PROFILE>

<PREFERENCE>
Preference-related information when available.
</PREFERENCE>

<UTILITY>
Theory-conditioned utility expression and evaluated choice
    probabilities.
</UTILITY>

<OPTIONS>
Valid choice labels and their descriptions.
</OPTIONS>

Instructions:
1. Ground the prediction in <KNOWLEDGE>.
2. Use <PROFILE> and <OPTIONS> to assess the plausibility of each
    choice.
3. Evaluate <UTILITY> as a theory-conditioned behavioral prior.
4. For PT, interpret the utility through reference-dependent gains and
    losses.
5. Output exactly one valid option label.

Debias:
Assume all options are equally likely before observing the provided
    evidence.
Do not rely on real-world prevalence.
```

```
[USR]
<KNOWLEDGE> {knowledge} </KNOWLEDGE>
<PROFILE> {profile} </PROFILE>
<PREFERENCE> {preference} </PREFERENCE>
<UTILITY>
  <EXPRESSION> {utility_expression} </EXPRESSION>
  <RESULT> {choice_probabilities} </RESULT>

  Prospect Theory information:
  V_pt,j = v(V_j - r)
  v(x) = x^alpha, if x >= 0
  v(x) = -lambda * (-x)^beta, if x < 0

  For each alternative:
  - raw utility V_j
  - deviation V_j - r
  - gain/loss status
  - PT-transformed probability
</UTILITY>
<OPTIONS> {options} </OPTIONS>

Return only the final option label.
```

Prompt A.5.2: PT Knowledge-Update Template

```
[SYS] You are an optimizer that improves the knowledge base of a
Prospect-Theory-guided decision agent.

In PT, alternatives are evaluated relative to a reference point.
A feature may move an alternative above or below the reference,
creating a gain or loss. Losses may be more psychologically salient
than gains, but numerical PT parameters are fitted separately.

Your task:
Revise the current knowledge base using prediction feedback.
The revised knowledge should remain domain-grounded and
reference-dependent.

Rules:
1. Revise no more than {MAX_EDITS} knowledge items.
2. Do not introduce external facts or dynamic information.
3. Do not invent numerical values for alpha, beta, lambda, or reference
   .
4. Prefer statements describing how features shift alternatives
   above or below a baseline.
5. Be specific about how feature interactions influence final choices.

Output format:
<IMPROVED_VARIABLE>
{revised_knowledge}
</IMPROVED_VARIABLE>

[USR]
Current knowledge:
{knowledge}

Prediction feedback:
{feedback}

Feature glossary:
{glossary}
```

```
Return only the tagged improved knowledge block.
```

Prompt A.5.3: PT Utility-Update Template

```
[SYS] You are an optimizer that improves a Prospect-Theory logit
choice model.

Each alternative has a raw utility score V_j. The model applies
reference-dependent valuation before softmax:

  V_pt,j = v(V_j - r)
  v(x) = x^alpha, if x >= 0
  v(x) = -lambda * (-x)^beta, if x < 0
  P(j) = softmax(V_pt,j)

Your task:
Revise the raw utility expressions V_j using:
1. textual feedback from prediction errors, and
2. numerical gradient feedback from the PT choice loss.

PT-specific constraints:
- Encourage cross-alternative spread around the reference point.
- Prefer expressions where alternatives can fall on different sides
  of the reference as features vary.
- Avoid identical formulas for all alternatives unless the task is
  truly symmetric.
- Keep gain/loss interpretation consistent with domain knowledge.
- Do not use trigonometric functions.

Output format:
<IMPROVED_VARIABLE>
<IMPROVED_VARIABLE_1>
{"alternative": "{option_1}", "utility": "{expression_1}"}
</IMPROVED_VARIABLE_1>
<IMPROVED_VARIABLE_2>
{"alternative": "{option_2}", "utility": "{expression_2}"}
</IMPROVED_VARIABLE_2>
<IMPROVED_VARIABLE_3>
{"alternative": "{option_3}", "utility": "{expression_3}"}
</IMPROVED_VARIABLE_3>
</IMPROVED_VARIABLE>

[USR]
Current utility expressions:
{current_utilities}

Numerical gradient feedback:
{gradient_feedback}

Textual prediction feedback:
{textual_feedback}

Task constraints:
{task_constraints}

Feature glossary:
{glossary}

Return only the tagged improved utility block.
```

The above prompts illustrate how PT-specific assumptions are injected into both forward reasoning and backward optimization. For other assumptions, the same prompt interface is used, but the <UTILITY> block and optimizer constraints are replaced by the corresponding choice map: MNL uses stable utility softmax, MXL uses heterogeneous random-coefficient choice probabilities, and PCL uses pairwise substitution probabilities. This design keeps the prompt structure consistent across theories while ensuring that each agent remains aligned with its own behavioral assumption.

A.6 TAL OPTIMIZATION SETTINGS

Setting	L2AO implementation
Outer iterations	16 iterations for each theory–group agent
Numerical fitting	Adam, 50 epochs, learning rate = 0.03
Stopping criterion	Fixed iteration budget; no gradient/loss convergence threshold
Validation checkpointing	Validation weighted-F1 evaluated after each outer update
Rollback	If validation F1 decreases, revert to the previous accepted state
Invalid utility expressions	Reject invalid proposals and retain the previous valid state

A.7 CONSENSUS DETAILS

Consensus threshold. We fix the consensus threshold to $\tau = 0.36$ across all datasets and experiments. The threshold is chosen to require a non-trivial confidence mass behind the consensus prediction, while still deferring genuinely weak-agreement cases to group–assumption alignment. Because our inference setting uses four theory-conditioned agents over three candidate labels, $\tau = 0.36$ is close to the one-third agreement level needed for a meaningful cross-theory signal. Empirically, nearby threshold values produce similar performance, indicating that L2AO is not sensitive to small changes around this setting.

Tie handling. When multiple labels receive the same highest vote count, the case is treated as non-consensus and passed to group–assumption alignment. Random tie-breaking is used only for the hard-vote ablation, with a fixed seed.

A.8 GROUP–ASSUMPTION ALIGNMENT DETAILS

This section provides the detailed construction of the group proxy and theory profile in CFEA. Throughout this section, we use the three-dimensional profile space

$$(\mathrm{per}, \mathrm{rob}, \mathrm{sem}),$$

where per denotes personalized preference sensitivity, rob denotes robustness under disagreement and uncertainty, and sem denotes gains from knowledge-guided reasoning beyond optimized utility.

We emphasize that the theory-derived profiles introduced below are *coarse inductive profiles* rather than exact theoretical quantities. Their role is to encode relative tendencies implied by different decision theories, while the final theory bias is corrected by dataset-level empirical evidence.

A.8.1 NORMALIZATION OPERATORS

For a scalar quantity u_g indexed by group g, we define the z-score normalization across groups as

$$z_g(u_g) = \frac{u_g - \mu(u)}{\sigma(u)}, \tag{11}$$

where $\mu(u)$ and $\sigma(u)$ are the mean and standard deviation over all groups in the current dataset. If $\sigma(u) = 0$, we set the normalized value to 0.

For a scalar quantity v_t indexed by theory t, we define min–max normalization across theories as

$$m_t(v_t) = \frac{v_t - \min_{t'} v_{t'}}{\max_{t'} v_{t'} - \min_{t'} v_{t'}}, \tag{12}$$

and set $m_t(v_t) = 0.5$ if all theories share the same value.

A.8.2 GROUP PROXY

For each demographic group g, we define the group proxy vector

$$\mathbf{h}_g = \left[h_g^{\text{per}},\, h_g^{\text{rob}},\, h_g^{\text{sem}}\right]. \tag{13}$$

Reject and disagreement subsets. Let $\hat{y}_{\text{cons}}(x)$ denote the consensus prediction and $\rho(x)$ its confidence-weighted agreement score as defined in the main text. We fix the consensus threshold to

$$\tau = 0.36. \tag{14}$$

A validation case is treated as a reject case when $\rho(x) < \tau$. To further identify severe disagreement, we define the high-divergence subset by

$$\mathcal{D}_g^{\text{div}} = \left\{x \in g : \big(\text{numUnique}(x) \geq 3\big) \;\vee\; \big(\text{voteGap}(x) \leq 1\big)\right\}, \tag{15}$$

and the reject-disagreement subset by

$$\mathcal{D}_g^{\text{rej}} = \{x \in g : \rho(x) < \tau\} \cap \mathcal{D}_g^{\text{div}}. \tag{16}$$

Personalized component. For each case x, let $a^\star(x)$ denote the correct agent with the highest confidence among all theories that predict the ground-truth label; if no theory is correct, the case contributes zero gain. We define

$$\Delta_g^{\text{sel}} = \frac{1}{|\mathcal{D}_g^{\text{rej}}|} \sum_{x \in \mathcal{D}_g^{\text{rej}}} \left(\mathbb{I}\Big[\hat{y}^{(a^\star(x))}(x) = y(x)\Big] - \mathbb{I}[\hat{y}_{\text{cons}}(x) = y(x)]\right), \tag{17}$$

which measures how much severe disagreement cases benefit from theory-specific correction over direct consensus.

We further define a minority-pressure statistic

$$\Delta_g^{\text{min}} = \frac{1}{|g|} \sum_{x \in g} \left(\bar{c}_{\text{minor}}(x) - \bar{c}_{\text{major}}(x)\right), \tag{18}$$

where $\bar{c}_{\text{minor}}(x)$ and $\bar{c}_{\text{major}}(x)$ denote the mean confidence of minority-vote and majority-vote theories on case x, respectively.

The personalized group bias is then defined as

$$h_g^{\text{per}} = \text{z}_\text{g}\big(\Delta_g^{\text{sel}}\big) + \text{z}_\text{g}\big(\Delta_g^{\text{min}}\big). \tag{19}$$

This component is intended to capture whether group g tends to require individualized correction beyond simple majority agreement. Groups with larger h_g^{per} are those for which minority-consistent theories or disagreement-specific corrections matter more.

Robustness component. We define a prediction-dispersion statistic

$$\Delta_g^{\text{var}} = \frac{1}{|g|} \sum_{x \in g} \text{Std}\big(c_a(x)\big)_{a \in \mathcal{A}}, \tag{20}$$

and a fragmentation statistic

$$\Delta_g^{\text{frag}} = \frac{1}{|g|} \sum_{x \in g} \left(\mathbb{I}[\rho(x) < \tau] + \mathbb{I}[\text{numUnique}(x) \geq 3]\right). \tag{21}$$

The robustness coordinate of the group proxy is then

$$h_g^{\text{rob}} = \text{z}_\text{g}\big(\Delta_g^{\text{var}}\big) + \text{z}_\text{g}\big(\Delta_g^{\text{frag}}\big). \tag{22}$$

This component measures how unstable or fragmented the assumption-conditioned agent pool appears for group g. A larger value indicates that the group is harder to model consistently and therefore requires stronger robustness-aware group–assumption alignment.

Semantic component. For each group g and behavioral assumption a, let $\mathrm{F1}_{g,a}^{\mathrm{util}}$ denote the optimized utility-only weighted F1, and let $\mathrm{F1}_{g,a}^{\mathrm{llm}}$ denote the corresponding guided LLM weighted F1. We define the group-level knowledge gain

$$\Delta_g^{\mathrm{know}} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \left(\mathrm{F1}_{g,a}^{\mathrm{llm}} - \mathrm{F1}_{g,a}^{\mathrm{util}} \right). \tag{23}$$

The semantic coordinate of the group proxy is then

$$h_g^{\mathrm{sem}} = z_{\mathrm{g}}\left(\Delta_g^{\mathrm{know}}\right). \tag{24}$$

This component quantifies how much group g benefits from knowledge-guided reasoning beyond the optimized utility representation. A larger value suggests that utility optimization alone is insufficient, and that guided LLM reasoning contributes substantially to final performance.

A.8.3 Theory Profile

For each assumption a, we define a dataset-level assumption profile vector

$$\mathbf{z}_a = \left[z_a^{\mathrm{per}},\, z_a^{\mathrm{rob}},\, z_a^{\mathrm{sem}} \right]. \tag{25}$$

Unlike $\mathbf{h}_g$, which is group-specific, $\mathbf{z}_a$ is dataset-specific and summarizes how assumption a behaves across the current task.

Theory-Derived Profile
We first assign each theory a coarse theory-derived profile

$$\mathbf{z}_a \in \mathbb{R}^3, \tag{26}$$

which reflects the relative inductive tendencies implied by the theory itself. For the four theories in our framework, we use

$$\mathbf{z}_{\mathrm{MNL}} = [0.2,\, 0.6,\, 0.2], \quad \mathbf{z}_{\mathrm{PCL}} = [0.5,\, 0.6,\, 0.4], \tag{27}$$

$$\mathbf{z}_{\mathrm{PT}} = [0.5,\, 0.5,\, 0.8], \quad \mathbf{z}_{\mathrm{MXL}} = [0.8,\, 0.5,\, 0.3]. \tag{28}$$

These values should be interpreted as coarse relative tendencies rather than exact theoretical truths.

MNL. MNL receives a low personalized score because it assumes homogeneous preferences and does not explicitly model latent heterogeneity. Its robustness score is set moderately high because its utility structure is simple, stable, and less prone to over-flexible fitting. Its semantic score is low because MNL itself does not naturally encode contextual or knowledge-driven reasoning beyond structured utility.

PCL. PCL receives a moderate personalized score because its pairwise-comparison structure captures richer local substitution patterns than MNL, making it more expressive in conflict-heavy decision settings. Its robustness score is also moderately high, reflecting its ability to relax simple independence assumptions while retaining structured comparability. Its semantic score is set to a moderate level, since pairwise comparative structure can support richer reasoning than plain MNL, though it remains less semantically oriented than PT.

PT. PT receives the highest semantic score because reference dependence, nonlinear valuation, and gain-loss asymmetry make it especially suitable for capturing behavior that deviates from optimized utility in psychologically meaningful ways. Its personalized and robustness scores are kept moderate: PT captures important behavioral distortions, but it is not a direct heterogeneity model and is more sensitive to framing assumptions than MNL.

MXL. MXL receives the highest personalized score because random coefficients explicitly model latent preference heterogeneity, making it the most natural theory for individualized behavioral variation. Its robustness score is kept moderate, since its additional flexibility can improve fit across diverse groups but does not guarantee stability in the same way as a simpler model. Its semantic score is set relatively low because its main strength lies in heterogeneity modeling rather than in contextual or knowledge-driven reasoning.

Empirical Profile

We next compute a dataset-level empirical profile $\mathbf{z}_{a,\mathrm{emp}}$ for each behavioral assumption a by aggregating validation behavior across groups. $\mathcal{D}$ denotes the validation set.

Personalized component. We define the assumption-level reject-case gain

$$\Delta_a^{\mathrm{rej}} = \frac{1}{|\mathcal{D}^{\mathrm{rej}}|} \sum_{x \in \mathcal{D}^{\mathrm{rej}}} \left(\mathbb{I}[\hat{y}_a(x) = y(x)] - \mathbb{I}[\hat{y}_{\mathrm{cons}}(x) = y(x)] \right), \tag{29}$$

where $\mathcal{D}^{\mathrm{rej}} = \{x : \rho(x) < \tau\}$ is the dataset-level reject subset.

We also define a minority-alignment score

$$\Delta_a^{\mathrm{min}} = \frac{1}{|\mathcal{D}^{\mathrm{min}}|} \sum_{x \in \mathcal{D}^{\mathrm{min}}} \mathbb{I}[\hat{y}_a(x) = y(x)], \tag{30}$$

where

$$\mathcal{D}^{\mathrm{min}} = \{x : \bar{c}_{\mathrm{minor}}(x) > \bar{c}_{\mathrm{major}}(x)\}. \tag{31}$$

Finally, let $\mathrm{F1}_{g,a}$ denote the weighted F1 of assumption a on group g. We define

$$\Delta_a^{\mathrm{var}} = \mathrm{Std}\big(\mathrm{F1}_{g,a}\big)_{g \in \mathcal{G}}. \tag{32}$$

The personalized empirical coordinate is

$$z_{a,\mathrm{emp}}^{\mathrm{per}} = m_a \big(\mathrm{z_a}(\Delta_a^{\mathrm{rej}}) + \mathrm{z_a}(\Delta_a^{\mathrm{min}}) + \mathrm{z_a}(\Delta_a^{\mathrm{var}}) \big), \tag{33}$$

where $\mathrm{z_a}(\cdot)$ denotes standardization across assumptions.

This component reflects whether assumption a is effective at correcting difficult disagreement cases and accommodating non-majoritarian patterns.

Robustness component. Let

$$\Delta_a^{\mathrm{conf}} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} c_a(x) \tag{34}$$

denote the average confidence of assumption a, and

$$\Delta_a^{\mathrm{f1var}} = \mathrm{Var}\big(\mathrm{F1}_{g,a}\big)_{g \in \mathcal{G}} \tag{35}$$

denote the variance of its group-level F1 across the dataset. The robustness empirical coordinate is

$$z_{a,\mathrm{emp}}^{\mathrm{rob}} = m_a \big(\mathrm{z_a}(\Delta_a^{\mathrm{conf}}) - \mathrm{z_a}(\Delta_a^{\mathrm{f1var}}) \big). \tag{36}$$

This component favors assumptions that remain confident while exhibiting stable performance across groups.

Semantic component. Let

$$\mathcal{D}^{\mathrm{low}} = \left\{ x : \max_{a \in \mathcal{A}} c_a(x) < \mathrm{median}_{x'} \max_{a \in \mathcal{A}} c_a(x') \right\} \tag{37}$$

denote the low-confidence subset. We define the low-confidence gain

$$\Delta_a^{\mathrm{low}} = \frac{1}{|\mathcal{D}^{\mathrm{low}}|} \sum_{x \in \mathcal{D}^{\mathrm{low}}} \left(\mathbb{I}[\hat{y}_a(x) = y(x)] - \mathbb{I}[\hat{y}_{\mathrm{cons}}(x) = y(x)] \right), \tag{38}$$

the average semantic advantage

$$\Delta_a^{\mathrm{sem}} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \left(\mathrm{F1}_{g,a}^{\mathrm{llm}} - \mathrm{F1}_{g,a}^{\mathrm{util}} \right), \tag{39}$$

and a high-semantic-win rate

$$\Delta_a^{\mathrm{win}} = \frac{1}{|\mathcal{G}^{\mathrm{high}}|} \sum_{g \in \mathcal{G}^{\mathrm{high}}} \mathbb{I}\left[a = \arg\max_{a' \in \mathcal{A}} \mathrm{F1}_{g,a'} \right], \tag{40}$$

where $\mathcal{G}^{\text{high}}$ denotes the subset of groups whose semantic group proxy exceeds the median.

The semantic empirical coordinate is then

$$z_{a,\text{emp}}^{\text{sem}} = m_a\big(\text{z}_\text{a}(\Delta_a^{\text{low}}) + \text{z}_\text{a}(\Delta_a^{\text{win}}) + \text{z}_\text{a}(\Delta_a^{\text{sem}})\big)\,. \tag{41}$$

This component is intended to capture how much assumption a benefits from knowledge-guided inference beyond optimized utility, especially in cases where the theory ensemble is uncertain.

A.8.4 Final Theory Profile

The final theory profile is obtained by blending the theory-derived and empirical profiles:

$$\mathbf{z}_a = \lambda \mathbf{z}_{a,\text{theory}} + (1-\lambda)\mathbf{z}_{a,\text{emp}}, \tag{42}$$

where

$$\lambda = 0.4. \tag{43}$$

This weighting reflects our design principle that theory-derived inductive structure should serve as a structured initialization, while the empirical profile should play the dominant role in adapting the theory profile to dataset-specific behavior.

B Dataset Details and Group Construction

We evaluate L2AO on five real-world survey-based decision-making tasks, covering transportation choice, vaccination behavior, employment status during COVID-19, and substance-use recency. All tasks are formulated as three-class classification problems. To model population heterogeneity, each dataset is partitioned into demographic subgroups defined by age and income-related attributes. Unless otherwise noted, each task uses a Cartesian partition of three age groups and three income groups, resulting in up to nine demographic groups per dataset. All models are trained, validated, and tested on fixed group-level splits.

B.1 Input Features

Table 3 summarizes the main feature categories used in each dataset. The selected features include both structured attributes, such as demographics, costs, time, and access constraints, and decision-relevant contextual or cognitive variables, such as perceived risk, trust, health status, and behavioral history.

B.2 Demographic Group Construction

For each dataset, demographic groups are constructed from age and income-related variables. The exact thresholds vary across datasets due to differences in survey design and variable availability.

SWM. SWM contains nine demographic groups constructed by crossing three age bins with three annual-income bins. The age groups are *under 40*, *40–65*, and *over 65*. The income groups are *under USD 50k*, *USD 50k–100k*, and *over USD 100k*.

VAC. VAC contains nine demographic groups constructed by crossing three age groups with three income-response categories. The age groups are *18–38*, *39–60*, and *61–87*. Income is represented by whether the respondent's income is *above the county median*, *below the county median*, or *refused / did not answer*. Different from the other datasets, VAC does not contain a continuous or tertile-based middle-income group; its income variable is defined relative to the county median, with an additional non-response category.

JOB. JOB contains nine demographic groups. The age groups are *below 35*, *35–65*, and *over 65*. The yearly household-income groups are *below USD 40k*, *USD 40k–100k*, and *over USD 100k*.

MAJ and ALC. MAJ and ALC use the same demographic grouping scheme. The age groups are *below 35*, *35–65*, and *over 65*. The yearly household-income groups are *below USD 50k*, *USD 50k–75k*, and *over USD 75k*.

Table 3: **Feature categories across datasets.** Each task includes structured demographic and contextual variables, together with task-specific behavioral signals.

Dataset	Feature categories
SWM (3; 35; 42)	Trip purpose, luggage, travel class, ticket type, car availability, travel pass ownership, age, income, gender, origin/destination indicators, and alternative-specific level-of-service attributes such as travel time, monetary cost, and headway for Train, Swissmetro, and Car.
VAC (25)	Vaccine-related beliefs, perceived COVID-19 threat, perceived vaccine safety and efficacy, trust in government and science, attention to vaccine information, age, gender, healthcare occupation indicators, education, and income-related indicators.
JOB (30)	Age, income, gender, education, self-rated health, chronic conditions, COVID symptoms, mental-health indicators, remote-work constraints, public-health restrictions, diagnosis status of self or household members, insurance, assistance, and test willingness.
MAJ (39)	Age, gender, education, income, age at first use, marijuana policy context, perceived availability, use urge, quitting difficulty, disorder-related indicators, risk-taking frequency, and attitudes toward marijuana use.
ALC (39)	Age, gender, education, income, high blood pressure, doctor visits, mental health, insurance and treatment access, workplace policy, binge-drinking history, risk-taking tendency, perceived harm of heavy drinking, alcohol screening items, and religiosity.

Table 4: **Dataset and demographic-group sample sizes.** Total denotes the number of observations used in each dataset after preprocessing and group construction. Test denotes the number of test observations, and Group size range reports the minimum and maximum number of observations across demographic groups.

Dataset	Total	Test	Group size range
ALC	967	283	76–114
JOB	876	244	85–109
MAJ	980	256	96–114
SWM	1008	279	97–114
VAC	1000	205	145–158*

*For VAC, the range is computed over the six age–income groups with observed above- or below-median income. Groups with refused income contain fewer observations and are excluded from the subsequent group-level analyses.

B.3 Class Balancing

For SWM, JOB, MAJ, and ALC, demographic cells are constructed using approximate class-balanced subsampling. When a cell contains more observations than the target size, samples are drawn to approximately balance the three outcome classes. If one class is underrepresented, the remaining slots are filled using available observations from the other classes. This procedure reduces severe label imbalance while preserving group-level demographic structure.

For VAC, the split is generated within each demographic cell using a fixed random train/validation/test partition. Unlike the other four tasks, VAC does not apply the same class-balanced subsampling procedure. Its class distribution therefore follows the empirical vaccination distribution within each demographic cell.

Table 4 summarizes the resulting dataset sizes and demographic-group sample sizes. For VAC, the reported group-size range corresponds to the six age–income groups with observed above- or below-median income, which are used in the group-level analyses.

C Baseline Implementation Details

All baselines are evaluated using the same train/validation/test splits as L2AO. For non-LLM baselines, we use fixed hyperparameter configurations rather than test-set tuning.

C.1 MLP (36)

Implementation: Training uses the default **Adam** solver. Maximum iterations: 600. Hyperparameters are chosen by validation weighted-F1 over a small grid: hidden layers in $\{(64), (128), (128, 64), (256, 128)\}$ and L2 strength $\alpha \in \{10^{-5}, 10^{-4}, 3 \times 10^{-4}\}$. When the training set has at least 16 samples, **early stopping** is enabled with `validation_fraction=0.1` (held out from training rows inside the classifier) and `n_iter_no_change=12`; otherwise early stopping is off. Missing features are imputed with the training median per column.

C.2 XGBOOST (4)

We train XGBoost as a multi-class classifier with probabilistic outputs. For all datasets, we use the multi-class objective `multi:softprob`, `mlogloss` as the evaluation metric, histogram-based tree construction, and all available CPU workers. No validation grid search is used.

We use:

$$\texttt{n_estimators} = 72, \quad \texttt{max_depth} = 4, \quad \texttt{learning_rate} = 0.048,$$
$$\texttt{min_child_weight} = 11, \quad \texttt{subsample} = 0.66, \quad \texttt{colsample_bytree} = 0.66,$$
$$\texttt{reg_lambda} = 5.5, \quad \texttt{reg_alpha} = 1.2.$$

C.3 PYSR (8) AND GPLEARN (38)

Since all tasks are formulated as three-class classification problems, both methods are implemented using a one-vs-rest strategy. For each class, we fit an independent binary symbolic model on the training fold. At test time, the final prediction is obtained by taking the class with the highest symbolic score.

PySR. We use:

$$\texttt{niterations} = 10, \quad \texttt{population_size} = 40, \quad \texttt{maxsize} = 9,$$
$$\texttt{parsimony} = 0.0025.$$

The operator set contains binary operators $\{+, -, \times, \div\}$ and no unary operators. We use `model_selection="best"` and `elementwise_loss="L2DistLoss()"`.

gplearn. We use:

$$\texttt{population_size} = 72, \quad \texttt{generations} = 5, \quad \texttt{max_samples} = 0.72,$$
$$\texttt{tournament_size} = 10, \quad \texttt{init_depth} = (1, 3), \quad \texttt{p_crossover} = 0.65.$$

The parsimony coefficient is fixed to `0.02`.

C.4 MNL, MXL, NL, PCL, PT, RRM

All six are estimated by **maximum likelihood** on the training indices of each group: the model defines choice probabilities $P(j \mid x)$ (multinomial logit, mixed logit simulation, nested logit, paired combinatorial logit, prospect-theory logit, or regret-based logit), and we minimize the **sum of negative log-likelihoods** over training observations. Optimization is **Adam** on all trainable parameters, defaulting to **50 epochs**, learning rate **0.03**, full-batch updates, **weight decay 0**, unless a wrapper overrides these. After training, we report metrics on the held-out **test** indices.

C.5 TEXTGRAD, ATHENA AND DUALMIND

TextGrad (46). Different from L2AO, this baseline only optimizes the textual system prompt. It does not include theory-specific utility functions, assumption-conditioned choice maps, group-level theory alignment, or consensus-based inference. For each dataset, the optimizable variable is the full system prompt used by the forward LLM predictor.

We run TextGrad for 6 outer optimization steps with batch size 3. During optimization, up to 50 validation prompts are used for checkpointing, and the best system prompt is selected according to

validation F1. If validation performance does not improve, the previous best prompt is retained. All runs use a fixed random seed of 42. When few-shot prompting is enabled, randomly sampled training exemplars are prepended to the prompt and kept fixed during optimization.

ATHENA (47). We adapt ATHENA as a two-stage baseline. Stage 1 discovers group-level symbolic utilities by prompting the LLM to generate alternative-specific utility expressions, fitting each candidate as a three-alternative MNL model, and using the best/worst candidates as feedback for iterative refinement. We use 30 proposal rounds, 5 candidates per round, top-2/bottom-1 feedback, GPT-4o-mini as the proposal model, and L-BFGS fitting with learning rate 1×10^{-4}, maximum 200 iterations, and 5 outer fitting epochs. Stage 2 optimizes a textual preference template initialized from the discovered symbolic utility. We use TextGrad with GPT-4o as the backward model, 5 optimization steps, and caps of 24 training and 16 validation prompts per group. ATHENA-MoR is evaluated with the same train/validation/test splits as L2AO. It differs from L2AO in that it learns one symbolic–textual pipeline per group, without multiple theory-conditioned agents, consensus-first inference, or group–assumption alignment.

DualMind (45). We evaluate DualMind under the same LLM configuration as L2AO. GPT-4o-mini is used as the forward prediction model, while GPT-4o is used as the backward model for generating textual feedback and updating the utility–knowledge representation. We use 16 optimization steps per group with validation-F1 checkpointing and rollback, retaining the previous best state when validation performance does not improve. Final predictions are generated by GPT-4o-mini using the optimized group-level utility and knowledge.

D Ablation at Component-level for CFEA

Table 5: **Ablation study at component level.** F1 and Acc are averaged across all datasets. P, R, and S denote the personal, robust, and semantic dimensions, respectively.

	Variant	F1 ↑	Acc ↑
Component	w/o P	0.678	0.689
	w/o R	0.677	0.682
	w/o S	0.676	0.686
	w/o P+R	0.674	0.680
	w/o P+S	0.676	0.688
	w/o R+S	0.672	0.682
	w/o Theory Profile	0.677	0.688
	w/o Empirical	0.669	0.680
Ours		**0.681**	**0.692**

Table 5 provides a component-level ablation of the group–assumption alignment module in CFEA. We remove each behavioral dimension from the group proxy and theory profile, including the personal (P), robust (R), and semantic (S) dimensions, and also test pairwise removals. The full model achieves the best performance, indicating that the three dimensions provide complementary signals for resolving non-consensus cases. Removing any single dimension leads to a small but consistent performance drop, while removing two dimensions causes larger degradation, particularly for *w/o R+S*. This suggests that alignment benefits from jointly modeling individualized variation, disagreement robustness, and knowledge-guided semantic needs.

We further ablate the two sources used to construct the theory profile. Removing the theory-derived profile slightly reduces performance, whereas removing the empirical validation profile causes a larger drop. This indicates that theory priors provide useful inductive structure, while dataset-specific empirical evidence plays the larger role in adapting assumption profiles to observed decision behavior.

Alternative aggregation mechanisms. To examine whether CFEA can be reduced to generic expert aggregation, we keep all TAL-trained theory-conditioned agents fixed and vary only the inference-time aggregation mechanism. As shown in Table 6, CFEA outperforms direct voting and weighting strategies and achieves performance comparable to learned MoE routing. While MoE-soft

Table 7: **Sensitivity to theory-pool composition.** Each variant removes one theory-conditioned agent while keeping the remaining pipeline unchanged.

Removed theory	F1	Accuracy
None (Full L2AO)	**0.681**	**0.692**
MXL	0.644	0.648
MNL	0.656	0.665
PCL	0.657	0.673
PT	0.641	0.641

attains a similar overall F1, CFEA retains an explicit and inspectable routing structure grounded in cross-theory consensus and group–theory compatibility, rather than relying on an additional learned gating model. This design makes the basis of each routing decision transparent and directly traceable to the behavioral assumptions represented by the theory-conditioned agents.

Table 6: **Comparison with alternative expert aggregation mechanisms.** All methods use the same TAL-trained theory-conditioned agents and differ only in how their outputs are combined at inference time.

Aggregation	F1 ↑	Acc ↑
Hard vote	0.656	0.663
Confidence-weighted	0.652	0.653
Validation-weighted	0.670	0.679
MoE-hard	0.659	0.656
MoE-soft	0.680	0.677
CFEA	**0.681**	**0.692**

Performance under theory disagreement. The advantage of CFEA becomes clearer on the cases where theory-conditioned agents disagree. Among 238 non-consensus cases, 134 are recoverable, meaning that at least one theory-conditioned agent predicts the correct outcome. CFEA recovers 56.7% of these cases, compared with 47.0% for validation-based group selection, 44.8% for MoE-soft, 34.3% for confidence-weighted aggregation, and 28.4% for hard voting. This indicates that CFEA mainly improves disagreement resolution rather than cases already resolved by cross-theory consensus.

Together, these results support CFEA as a hybrid alignment mechanism that combines theory-grounded behavioral priors, validation-derived group evidence, and consensus-aware expert coordination.

E Robustness and Sensitivity Analysis

We further examine whether L2AO is sensitive to the theory pool, demographic-group construction, or data split. Across these perturbations, the main performance advantage remains stable, suggesting that the reported results are not tied to a particular theory composition, manually chosen group boundary, or train–validation–test split. Because these analyses require retraining multiple theory-conditioned agents across demographic groups, repeating the full evaluation on all five datasets would incur substantial LLM optimization cost. We therefore conduct the detailed sensitivity analyses on two complementary datasets, SWM and VAC. They provide substantially different decision environments for testing sensitivity to demographic construction and data resampling.

Sensitivity to theory-pool composition. We conduct a leave-one-theory-out analysis while keeping the remaining training and inference pipeline unchanged. As shown in Table 7, removing any individual theory reduces performance, but no single removal causes the framework to collapse. This supports the role of the four theories as complementary behavioral assumptions rather than dependence on one dominant theory.

Sensitivity to demographic-group boundaries. We perturb the age boundaries in SWM and VAC while leaving the remaining pipeline unchanged. The alternative partitions move 13.4% and 15.0% of observations to different demographic groups, respectively. As shown in Table 8, the resulting F1 changes are small (0.014 for SWM and 0.002 for VAC), indicating that L2AO is not strongly dependent on the exact demographic thresholds.

Table 8: **Sensitivity to demographic-group boundaries.** Alternative age partitions substantially change group membership while preserving the remaining experimental protocol.

Dataset	Age partition	F1	Accuracy
SWM	Original: < 40 / 40–65 / > 65	**0.694**	**0.695**
SWM	Alternative: < 40 / 40–55 / > 55	0.680	0.687
VAC	Original: < 39 / 39–60 / > 60	**0.669**	0.649
VAC	Alternative: < 36 / 36–55 / > 55	0.667	**0.669**

Robustness to data splits. We additionally regenerate the train–validation–test splits for SWM and VAC, replacing 72–83% of the original test samples. Table 9 reports L2AO together with representative PT and knowledge-enhanced vanilla-LLM baselines. L2AO maintains similar performance across independently generated splits, with lower variability than PT on both datasets.

Table 9: **Robustness to independently regenerated data splits.** Mean and standard deviation are computed over the original split and two independently regenerated splits.

Dataset	**Split**	**L2AO**	**PT**	**LLM-V+K**
VAC	Original	0.669	0.616	0.351
	Split 1	0.674	0.623	0.432
	Split 2	0.629	0.547	0.324
	Mean $\pm$ s.d.	**0.657** $\pm$ **0.025**	0.595 ± 0.042	0.369 ± 0.056
SWM	Original	0.694	0.464	0.491
	Split 1	0.686	0.516	0.471
	Split 2	0.676	0.638	0.488
	Mean $\pm$ s.d.	**0.685** $\pm$ **0.009**	0.540 ± 0.089	0.484 ± 0.011

Overall, these analyses show that L2AO remains effective under changes to the available theory set, demographic-group construction, and data partition, supporting the robustness of the main results to these design choices.

F Compute Resources

Our experiments mainly rely on API-based LLM calls rather than local GPU training. All experimental scripts are executed on a Mac Studio workstation, which is used for data preprocessing, prompt construction, API orchestration, result parsing, and metric computation. The main computational cost comes from LLM-guided optimization. For each dataset, we train theory-conditioned agents under four behavioral assumptions: MNL, MXL, PT, and PCL. Each single-theory agent is optimized at the group level for 16 steps over up to nine demographic groups, with rollback when validation F1 decreases. Training one single-theory agent on one dataset takes approximately 35 minutes in our setup. For each dataset, the four single-theory agents are trained independently, and we repeat the single-theory agent training three times to compute averaged results. The API usage for one dataset-level single-theory agent training is approximately 1×10^8 to 1.5×10^8 tokens, depending on dataset size and prompt/response length.

We use GPT-4o-mini for forward reasoning and final prediction, and GPT-4o only for TextGrad-based backward optimization. Thus, the total execution cost scales primarily with the number of datasets, behavioral assumptions, and optimization steps, rather than with local hardware capacity. Formally, the optimization cost scales as $O(|\mathcal{D}||\mathcal{G}||\mathcal{T}|S)$, where $|\mathcal{D}| = 5$ is the number of datasets, $|\mathcal{T}| = 4$ is the number of behavioral assumptions, $|\mathcal{G}| = 9$ is the number of demographic groups per dataset, and $S = 16$ is the number of optimization steps.