Personalized Decision Modeling: Utility Optimization or Textualized-Symbolic Reasoning

Yibo Zhao
Department of Civil and Systems Engineering
Johns Hopkins University

Yang Zhao
Department of Civil and Systems Engineering
Johns Hopkins University

Hongru Du [*] [†]
Department of Systems and Information Engineering
University of Virginia
hongrudu@virginia.edu

Hao Frank Yang [†]
Department of Civil and Systems Engineering
Johns Hopkins Data Science and AI Institute
Johns Hopkins University
haofrankyang@jhu.edu

Abstract

Decision-making models for individuals, particularly in high-stakes scenarios like vaccine uptake, often diverge from population optimal predictions. This gap arises from the uniqueness of the individual decision-making process, shaped by numerical attributes (e.g., cost, time) and linguistic influences (e.g., personal preferences and constraints). Developing upon Utility Theory and leveraging the textual-reasoning capabilities of Large Language Models (LLMs), this paper proposes an Adaptive Textual-symbolic Human-centric Reasoning framework (**ATHENA**) to address the optimal information integration. ATHENA uniquely integrates two stages: First, it discovers robust, group-level symbolic utility functions via LLM-augmented symbolic discovery; Second, it implements individual-level semantic adaptation, creating personalized semantic templates guided by the optimal utility to model personalized choices. Validated on real-world travel mode and vaccine choice tasks, ATHENA consistently outperforms utility-based, machine learning, and other LLM-based models, lifting F1 score by at least 6.5% over the strongest cutting-edge models. Further, ablation studies confirm that both stages of ATHENA are critical and complementary, as removing either clearly degrades overall predictive performance. By organically integrating symbolic utility modeling and semantic adaptation, ATHENA provides a new scheme for modeling human-centric decisions. The project page can be found at https://yibozh.github.io/Athena.

1 Introduction

Consider the widely debated *vaccine dilemma* [1], from a population-level perspective aimed at optimizing collective well-being (e.g., achieving herd immunity at minimal societal cost), models would invariably predict near-universal vaccine adoption. However, this population optimum consistently

[*]This work was completed while Hongru Du was at Johns Hopkins University.
[†]Correspondence to: Hongru Du and Hao Frank Yang.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

fails to predict actual individual behavior. The reality is a broad spectrum of personal choices, because each individual undertakes their own unique *cognitive calculus*: balancing a vaccine's perceived efficacy and protection against its perceived risks [2], all filtered through their personal beliefs, constraints, and even considerations of potential "free-riding" on the immunity of others [3]. This divergence between the theoretical population optimum and observed individual actions highlights a critical limitation: **models designed to maximize collective outcomes do not adequately explain or predict individuals' choices.** Instead, individual choices are profoundly shaped by who we are, when the decision is made, and the unique constraints we face. This internal *'cognitive calculus'*, unique to each individual and situation, presents a profound challenge for human decision modeling.

For decades, researchers have attempted to model this *'cognitive calculus'* with Utility Theory [4–6], which assumes individuals select options that maximize expected gain. Operationally, this involves defining a parametric utility function, denoted as $f : \mathcal{X} \rightarrow \mathbb{R}$, that maps a vector of structured attributes $\mathcal{X}$ (e.g., monetary cost and time) for each option to a scalar utility score. Such pre-defined and explicit specifications of f are the basis for classic discrete choice models [7], which have been widely adopted across economics [8–10], transportation [7, 11–13], and public policy [14–16]. In these models, the utility scores derived from f for each available option are used to probabilistically determine the likelihood of an individual selecting a particular option. However, even these utility-based models encounter fundamental barriers when attempting to capture the full depth of human decision-making. Real-world human decisions, as seen in the *vaccine dilemma*, frequently deviate from these mathematical formulations. Individuals exhibit behaviors that appear inconsistent or irrational [17, 18], yet these are often driven by subjective feelings and personal experiences. Such deviations reflect that traditional models, with their reliance on pre-specified functions, struggle to capture personalized decisions [19]. The clue for deciphering this deviation is covered within individual attributes, some of which are structured and quantifiable, while others are unstructured and semantic (e.g., individual preference and constraints).

Addressing these unstructured and semantic dimensions, which are pivotal for capturing personalized decision mechanisms, calls for new modeling paradigms. LLMs with their strong textual-reasoning capability offer a clear advance [20], providing new mechanisms for identifying the utility function f and for integrating semantic individual context directly into the decision modeling process. Specifically, LLMs enhance our ability to model human choice by: a) Guiding the discovery of more accurate and robust parametric forms for f. Through LLM-augmented symbolic regression [21–23], it becomes feasible to identify data-driven mathematical structures that capture underlying group-level choice patterns more effectively than pre-specified forms. b) Enabling the infusion of individual-level textual information into the human decision modeling [24–26]. By encoding personal preferences, constraints, and narratives, LLMs allow each decision to reflect the nuanced motivations and situational factors that traditional numeric features alone cannot convey.

This paper introduces an **Adaptive Textual-symbolic Human-centric Reasoning framework (ATHENA)**. ATHENA achieves personalized decision modeling by uniquely integrating two sequentially structured steps: First, at the group level, it focuses on discovering robust, symbolic utility functions. Second, it implements individual-level, LLM-powered semantic adaptation guided by optimal utility functions discovered in previous steps. The outcome is a customized semantic template for each person, specifically designed to empower an LLM to model their choices by incorporating their unique preferences and constraints. We empirically validate ATHENA on two real-world human decision-making tasks: travel mode choice and vaccine uptake decisions. The model consistently outperforms traditional utility-based, machine learning, and LLM-based models, with at least 6.5% improvement in F1 score. Further ablation experiments reveal that removing either the group-level symbolic utility search or the individual semantic adapter lowers performance by at least 18%, underscoring the merit of the full ATHENA framework.

2 Related Work

Utility-based Decision-Making Models. Initial explorations into this complex domain were predominantly by utility-based models [27–30]. These methods aim to capture human behavior within explicit mathematical functions, formulated from empirical data. Established methodologies such as Discrete Choice Models (DCMs) have been widely used, attributable to their interpretability and robust statistical underpinnings [31, 10, 32–37]. While offering tractability, they may also limit the ability to fully capture complex non-linear patterns and diverse preferences in modern high-dimensional data.

Machine Learning–Driven Decision-Making Models. ML-driven decision-making models aim to directly learn from rich, diverse features. Tree-based ensemble methods, including Random Forests [38], Gradient-Boosting Trees [39], XGBoost [40, 41], and LightGBM [42], alongside neural network architectures [43–45], exhibited a notable proficiency in fitting complex, non-additive interaction effects. These models effectively integrated large-scale data, but their decision-making processes often lacked transparency despite strong predictive performance. Efforts to enhance transparency via post-hoc explanation frameworks, for instance, SHAP [46–48] and Integrated Gradients [49, 50], have provided some insights for human behavior. A persistent challenge is these models' vulnerability to distribution shifts, lack of transparency, and limited ability to provide symbolic, interpretable insights needed for personalized utility reasoning.

Symbolic Regression with LLMs. Classical symbolic regression (SR) typically uses genetic programming to evolve populations of candidate equations via stochastic mutation and crossover [51–53]. The goal is to find formulas that balance simplicity, generalizability, and human interpretability [54–56]. The recent rise of LLMs has revitalized symbolic regression, enabling new possibilities in scientific discovery when combined with advanced evolutionary algorithms [57]. For example, LLM-SR integrates LLM with evolutionary symbolic regression by treating equations as executable programs. It leverages LLMs' scientific prior knowledge and code generation abilities to iteratively generate, refine, and optimize equation skeletons [22]. LASR integrated LLM-driven abstract textual concepts within evolutionary frameworks, achieving notable performance enhancements on benchmarks, such as the Feynman equation set [58]. The DiSciPLE framework extended these contributions by emphasizing the interpretability and reliability of generated scientific hypotheses, incorporating critical evaluation and simplification to ensure hypotheses are both scientifically rigorous and computationally efficient [59].

LLM-based Decision-Making Models. The advent of LLMs has offered a new opportunity, establishing these models as human-like reasoning engines [60–65]. Techniques such as instruction tuning [66–69], chain-of-thought reasoning [70–74] are elevating LLMs move beyond basic text generation to handle more complex tasks involving step-by-step reasoning and symbolic or numerical problem-solving [75–81]. Within the specific domain of personalized decision making, preliminary findings suggest that zero-shot and few-shot prompting strategies can enhance the behavioral alignment of LLMs [82, 83]. Because a model's knowledge is inherited from generic pre-training priors, its reasoning defaults to universally salient factors – e.g., cost and time in travel mode choice – while overlooking personal preferences such as rail-pass loyalty or transfer aversion, thereby introducing systematic bias [84]. Techniques like persona loading partially mitigate this gap by conditioning responses on inferred preference structures [82, 85]. Beyond basic prompting, decision-centric systems add explicit structure to improve reliability and transparency: *DeLLMa* enumerates plausible states, elicits utilities via pairwise comparisons, and maximizes expected utility; *STRUX* distills inputs into fact tables with self-reflective evidence; *OptiGuide* compiles natural-language "what-if" queries into optimization code and invokes solvers; *Agent-Driver* coordinates tool calls, commonsense/experience memory, and chain-of-thought planning; *Personalized Oncology* evaluations show chat-LLMs still trail experts, motivating structured, evidence-grounded pipelines [86–90]. Nevertheless, many deployments still treat LLMs as opaque scoring mechanisms, falling short of fully recovering explicit, personalized utility logic.

3 Methods

We consider a classic discrete choice problem, where an individual i faces a finite set of choices $\mathcal{J} = \{1, 2, \ldots, J\}$, where $J = |\mathcal{J}|$. The decision-making process assumes individuals select the option j that maximizes their utility. Each individual's observed choice behavior is represented by a dataset $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$. For each observation i, $X_i = \{x_{ij}\}_{j=1}^{J}$ represents the set of feature vectors, where $x_{ij} \in \mathbb{R}^K$ captures the features for choice j and $y_i \in \mathcal{J}$ denotes the observed choice.

The standard approach to modeling discrete choices is the Random Utility Maximization (RUM) framework [91]. It assumes the latent utility for each alternative j is described by $U_{ij} = f(X_i, j) + \epsilon_{ij}$, where $f(X_i, j)$ is systematic component of utility and ϵ_{ij} is the random error. Assuming ϵ_{ij} are independently and identically drawn (i.i.d.) and follow a Type I Extreme Value distribution [92], the probability of individual i choosing alternative j is given by:

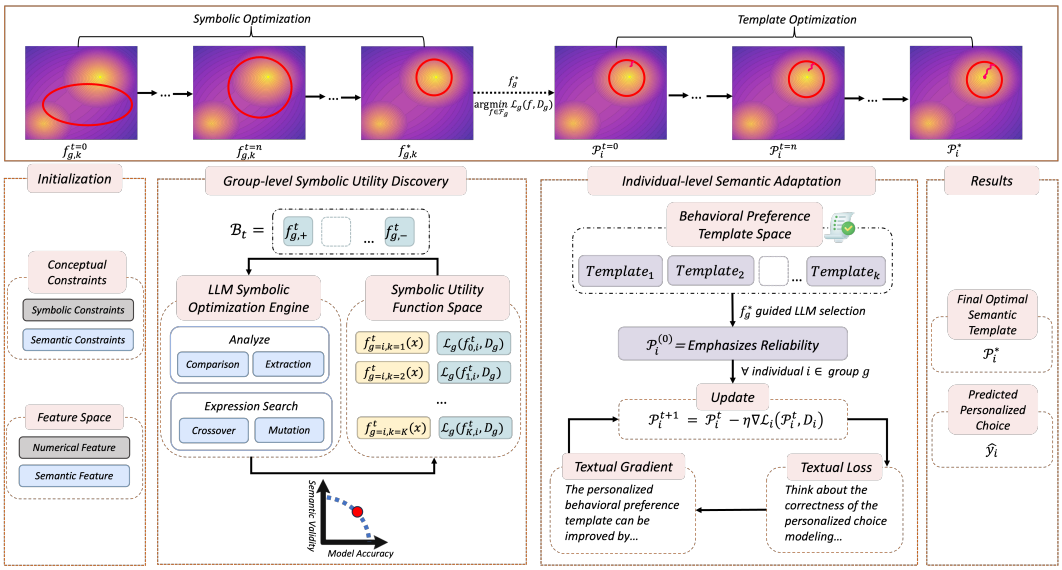


Figure 1: **Overview of the proposed ATHENA framework.** *Group-level symbolic utility discovery*: Symbolic & semantic constraints library feed an LLM-driven symbolic-optimization engine that iteratively proposes candidate utility functions, scores them with loss $\mathcal{L}_g$, and prunes the search via analysis, crossover, and mutation. Red rings in the contour maps illustrate how the feasible solution space shrinks across iterations until the optimal formula f_g^* is selected. *Individual-level semantic adaptation*: The optimal group utility f_g^* seeds a personalized template space. For each individual i, TextGrad computes textual gradients of an individual loss and updates the template $\mathcal{P}_i^t$ into a more personalized decision rule $\mathcal{P}_i^{t+1}$. Finally, the optimal $\mathcal{P}_i^*$ is used to predict personal decisions.

$$P(y_i = j \mid X_i) = \frac{e^{f(X_i, j; \theta)}}{\sum_{k \in \mathcal{J}} e^{f(X_i, k; \theta)}}. \tag{1}$$

A central challenge lies in specifying the systematic utility component f. Traditional applications of RUM often rely on pre-specified functional forms for f using domain expertise and observed data [93]. This approach may result in a suboptimal representation of the true decision mechanism, while also neglecting individual heterogeneity in choices. Furthermore, traditional decision-making models are not designed to incorporate non-structured semantic information. To address these limitations, we introduce ATHENA for personalized decision modeling designed to identify suitable utility function representations while simultaneously capturing individual-specific preferences. As shown in Fig. 1, ATHENA structures the decision modeling process into two sequential stages:

1. **Group-Level Symbolic Utility Discovery:** This initial stage focuses on identifying optimal symbolic utility components that capture common decision patterns within distinct demographic groups. The discovery is achieved through a feedback-informed symbolic discovery process powered by LLMs.

2. **Individual-Level Semantic Adaptation:** Then, the optimal group-level utility functions serve as guidance for the LLM-driven optimization of personalized semantic templates. This adaptation process is designed to incorporate individual-specific preferences and constraints, leveraging the rich semantic reasoning capabilities of LLMs.

3.1 Group-Level Symbolic Utility Discovery

The first stage aims to discover an optimal parametric utility function, denoted as f_g^*, for each demographic group $g \in \mathcal{G}$. This function f_g^* should be constructible from symbolic building blocks and optimally explain the group's choice behavior. Following the choice probability defined earlier

Eq. (1), the objective is to find the optimal f_g^* and its associated parameters θ_g^* such that:

$$(\theta_g^*, f_g^*) = \underset{f,\theta}{arg\,min}\ \mathcal{L}_g(f(X_i, y_i; \theta); \mathcal{D}_g) = -\sum_{(X_i,y_i)\in\mathcal{D}_g} \log\left(\frac{e^{f(X_i,y_i;\theta)}}{\sum_{k\in\mathcal{J}} e^{f(X_i,k;\theta)}}\right). \tag{2}$$

To automate the symbolic utility discovery of f_g^*, we design an iterative, feedback-informed generation process powered by LLMs. To effectively guide the automated discovery of utility functions, we constructed two foundational libraries: a domain knowledge concept library ($\mathcal{C}$) and a symbolic library ($\mathcal{S}$). The library $\mathcal{C}$, developed based on input from domain experts, covers high-level conceptual knowledge about domain-specific human behavior. The library $\mathcal{S}$ provides the fundamental syntactic building blocks needed for constructing all candidate utility expressions.

Inspired by evolutionary algorithms [94], the core discovery process proceeds iteratively. In each iteration t for each demographic group g, the LLM samples a set of K candidate symbolic utility functions, $\{f_{g,k}^t\}_{k=1}^K$. This sampling is performed from the LLM's learned distribution ϕ [20, 95], conditioned on the group profile g, the domain concept $\mathcal{C}$, the available symbolic building block $\mathcal{S}$, and a feedback $\mathcal{B}^{t-1}$ from preceding iteration:

$$\{f_{g,k}^t\}_{k=1}^K \sim \phi(\cdot|g, \mathcal{C}, \mathcal{S}, \mathcal{B}^{t-1}) \tag{3}$$

The feedback $\mathcal{B}$ is essential in refining the LLM's sampling strategy. Specifically, $\mathcal{B}^t$ is constructed at the end of each iteration t and comprises the best-performing and worst-performing candidate functions from that iteration:

$$\mathcal{B}^t = \{f_{g,+}^t, f_{g,-}^t\}, \tag{4}$$

where $f_{g,+}^t = \underset{k\in K}{arg\,min}\ \mathcal{L}_g(f_{g,k}^t, \mathcal{D}_g)$ and $f_{g,-}^t = \underset{k\in K}{arg\,max}\ \mathcal{L}_g(f_{g,k}^t, \mathcal{D}_g)$. Here $\mathcal{L}_g$ is the group-level loss function, with a similar format as Eq. (2). This feedback $\mathcal{B}^t$ is used to refine the LLM's sampling distribution ϕ through stochastic *mutation* or *crossover* [51–53], pushing the generation towards more promising types of functions. The iterative discovery process for group g is considered to have converged at iteration T if the absolute difference in the loss of the best-performing function from the current iteration t and that of the previous iteration $t-1$ falls below a predefined threshold δ:

$$|\mathcal{L}_g(f_{g,+}^t, \mathcal{D}_g) - \mathcal{L}_g(f_{g,+}^{t-1}, \mathcal{D}_g)| < \delta. \tag{5}$$

Upon convergence at iteration at T, the optimal group-level symbolic utility function (f_g^*) is determined as the function that achieved the minimum loss across all generated candidate functions throughout the entire iterative process:

$$f_g^* = \arg\min_{f\in\mathcal{F}_g} \mathcal{L}_g(f, D_g), \tag{6}$$

where $\mathcal{F}_g = \bigcup_{t=1}^T \{f_{g,k}^t\}_{k=1}^K$ and T is the iteration at which convergence occurred. This discovered function f_g^*, along with its fitted parameters θ_g^*, serves as the learned representation of the systematic utility for group g.

3.2 Individual-Level Semantic Adaptation

Following the determination of the group-level optimal symbolic utility functions f_g^*, the framework transitions to the second stage, leveraging an LLM conditioned on f_g^* to model individual choice behavior more accurately. While f_g^* captures the central tendencies of utility for group g, significant intra-group heterogeneity often persists. To account for this, we introduce an individual-level adaptation stage to personalize the utility representation by generating and refining an individual-specific semantic template.

For each individual $i \in g$, the initial semantic template, denoted as $\mathcal{P}_i^0$ is generated by the LLM (ϕ). The generation of the initial semantic template is represented as a sampling process from the LLM's distribution: $\mathcal{P}_i^0 \sim \phi(\cdot|f_g^*, i, \mathcal{C})$. In this formulation, ϕ conditions on the optimal group-level symbolic function f_g^*, the specific individual context i, and the high-level domain concepts from $\mathcal{C}$ to generate $\mathcal{P}_i^0$. This initial template $\mathcal{P}i^0$ is a semantic representation that is designed to be adaptable in

subsequent optimization steps. Then, the semantic template $\mathcal{P}_i^0$ undergoes an iterative refinement process for each individual i. This optimization is driven by TextGrad [96], which optimizes the template based on the individual's specific data $\mathcal{D}_i = (X_i, y_i)$. The update rule is given by:

$$\mathcal{P}_i^{t+1} \leftarrow \mathcal{P}_i^t - \eta \nabla \mathcal{L}_i(\mathcal{P}_i^t, D_i). \tag{7}$$

The term $\nabla \mathcal{L}_i(\mathcal{P}_i^t, D_i)$ represents the "textual gradient" of the loss function with respect to the semantic template $\mathcal{P}_i^t$. Since $\mathcal{P}_i^t$ is the textual template, this gradient is not a vector of partial derivatives in the mathematical sense. Instead, it indicates the direction and nature of textual modifications to $\mathcal{P}_i^t$ that would lead to the most improvement in loss. This iterative refinement process continues until a maximum number of iterations T' is reached. Then the final optimal semantic template for individual i, denoted as $\mathcal{P}_i^*$, is determined. The predicted personalized choice $\hat{y}_i$ is then represented as sampling from the LLM's output distribution:

$$\hat{y}_i \sim \phi(\underbrace{\mathcal{P}_i^*, X_i}_{\text{Semantic Adaptation}} \mid \overbrace{f_g^*(X_i; \theta_g^*)}^{\text{Symbolic Utility Discovery}}) \tag{8}$$

The overall procedure of ATHENA is summarized in Algorithm 1.

Algorithm 1 ATHENA Optimization Flow

Require: Demographic group g, dataset $\mathcal{D}_g$, domain concept $\mathcal{C}$, symbolic building block $\mathcal{S}$

Initialize $\mathcal{B}_0 \leftarrow$ `None`

// Stage 1: Group-Level Symbolic Utility Discovery

for $t = 1$ to T **do**

  Sample symbolic utility functions $\{f_{g,k}^t\}_{k=1}^K \sim \phi(\cdot \mid g, \mathcal{C}, \mathcal{S}, \mathcal{B}^{t-1})$

  Update $\mathcal{B}^t \leftarrow \{f_{g,+}^t, f_{g,-}^t\}$ using Eq. (4)

  Select best function $f_g^* \leftarrow \arg\min_{f \in \mathcal{F}_g} \mathcal{L}_g(f, \mathcal{D}_g)$

  if stopping condition in Eq. (5) is met **then**

    break

  end if

end for

// Stage 2: Individual-Level Semantic Adaptation

for each individual $i \in g$ **do**

  Initialize semantic template $\mathcal{P}_i^0 \sim \phi(\cdot \mid f_g^*, i, \mathcal{C})$

  for $t = 1$ to T' **do**

    Update $\mathcal{P}_i^{t+1} \leftarrow \mathcal{P}_i^t - \eta \nabla \mathcal{L}_i(\mathcal{P}_i^t, \mathcal{D}_i)$ using Eq. (7)

  end for

end for

return $\{\mathcal{P}_i^*\}_{i \in g}$, predict decisions using Eq. (8).

4 Experiments

This section empirically validates the value of ATHENA, demonstrating its overall effectiveness in personalized decision-making and its robust capability to apply across diverse application domains. We break down our experimental findings to specifically showcase the distinct value added by each core component of the ATHENA framework: 1) group-level symbolic utility discovery and 2) personalized semantic template adaptation. Fig. 2 illustrates the full pipeline using the travel-mode choice as an example.

4.1 Experimental Setup

Datasets. To test ATHENA's ability to generalize across different domains and to adapt to individual preferences, we selected two real-world tasks that reflect fundamentally different personalized decision scenarios: daily transportation choices and public health decisions. **(1) Swissmetro Transportation Choice (*Swissmetro*)**: is a widely used benchmark in travel mode choice modeling [97–101]. Each record details a trip between major Swiss cities and includes both traveler characteristics (e.g., income, age) and alternative-specific attributes (e.g., travel time, cost). The dataset has a potential choice set of three transportation modes: *Train*, *Car*, and *Metro*. **(2) COVID-19 Vaccination Choice (*Vaccine*)**: This dataset is derived from a large-scale international survey, conducted across multiple countries [102]. The survey was designed to understand factors influencing COVID-19 vaccine uptake and attitudes. For each participant, it captures demographics, prior beliefs about the vaccine, and their self-reported vaccination status. The modeled choices based on this information include: *Unvaccinated*, *Vaccinated initial doses, no booster*, and *Vaccinated initial doses plus booster*.

Experiment Configurations. To maintain a reasonable budget for the template-adaptation stage, we restricted the experimental sample to a representative subset of each dataset. Specifically, we used: *(1)* ***Swissmetro***: 500 travelers, two trip records per person; *(2)* ***Vaccine***: 300 respondents, one survey record per person. Within each dataset, we first identified key demographic dimensions (gender,

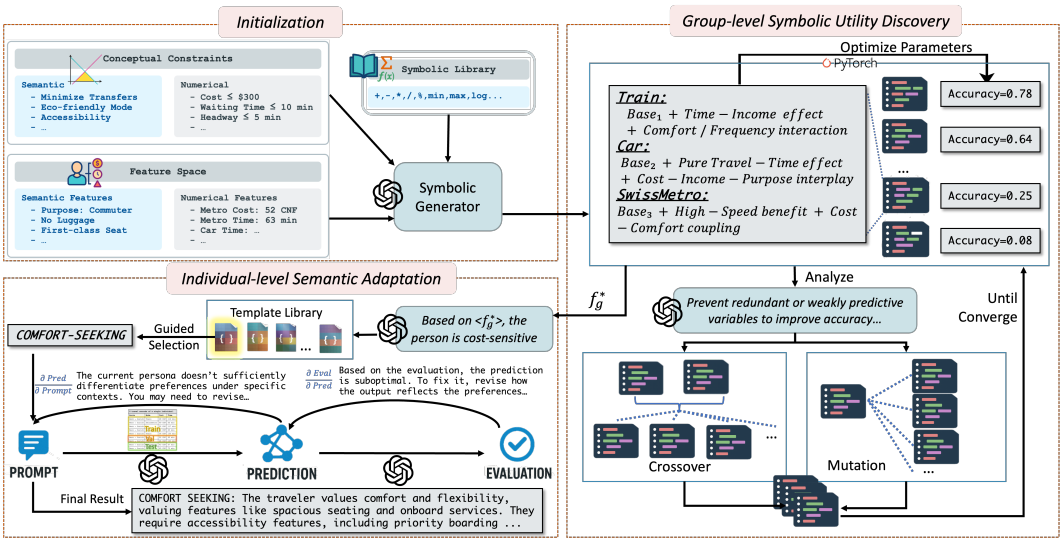


Figure 2: **ATHENA pipeline applied to a travel–mode choice example.** Here we use ***Swissmetro*** as an example to illustrate ATHENA framework. The *Initialization* panel encodes conceptual constraints, a mixed semantic–numerical feature space, and a symbolic library of operations. In *Group-level symbolic optimization*, an LLM samples, scores, and prunes candidate utility expressions for each alternative to produce compact formulas $\{f_g^*\}$ that best explain group behavior. In *Individual semantic adaptation*, each f_g^* seeds a group-specific prompt template $\mathcal{P}_i^0$, which is refined to a personalized template via TextGrad to capture individual heterogeneity ($\mathcal{P}_i^0 \rightarrow \mathcal{P}_i^*$).

age, and income), then sampled approximately balanced subsets across these strata from the full dataset. This ensures *(i)* comparable class priors between training and test splits, and *(ii)* that no demographic group dominates the symbolic-utility discovery process. The predefined demographic grouping follows established practice in choice modeling, supports interpretability, and improves robustness by avoiding the complexity and data requirements of latent clustering methods [103–105].

Evaluation metrics. We report Accuracy, F1, AUC, and Cross-Entropy (CE). CE is included because ATHENA produces probabilistic predictions over choices. A lower CE means the model assigns higher probabilities to actual choices, while F1 and AUC capture classification performance; together, they provide complementary views on accuracy and calibration.

Models and baselines. Both stages of ATHENA, symbolic-utility discovery and individual semantic adaptation, run on the `gpt-4o-mini-2024-07-18` and `gemini-2.0-flash`. To evaluate its performance, we contrasted ATHENA with three baseline groups. *(i)* LLM-based methods: a plain zero-shot method [106, 107], a zero-shot chain-of-thought method [106], a five-example few-shot method [108, 109], and TextGrad tuning [96]. *(ii)* Classical discrete-choice models: Multinomial Logit (MNL) [110], Conditional Logit (CLogit) [111], and Latent-Class MNL [112]. *(iii)* Standard machine-learning classifiers: logistic regression, random forest, XGBoost [113], a shallow two-layer MLP [114], TabNet for tabular data [115], and a fine-tuned BERT classifier [116]. This spectrum ranges from end-to-end language-model reasoning through discrete choice models to conventional predictive learners, providing a balanced reference for unique modeling capabilities.

4.2 Overall Performance Analysis

Performance and insights. As shown in Table 1, on the *Swissmetro* mode choice task, ATHENA with GPT-4o-mini notably outperforms evaluated baselines across Accuracy (Acc), F1-score (F1), and AUC. Over the strongest baseline, it achieves gains of at least 6% in Acc and 6.5% in F1, respectively. Similar improvements are noted on the *Vaccine* dataset. Notably, our proposed method exhibits higher Cross-Entropy (CE) compared to baselines such as XGBoost. We attribute this to the inherent design of ATHENA, which produces more conservative probability distributions rather than extreme certainties. Specifically, unlike models that might predict a choice with $> 90\%$ confidence, ATHENA's framework is less prone to such high probabilities. This characteristic may better reflect the

Table 1: Performance comparison of LLM-based, classical choice, and machine learning methods on the three-class Swissmetro and three-class COVID-19 Vaccine choice tasks.

	Method	LLM Model	Swissmetro				Vaccine			
			Acc.↑	F1.↑	CE.↓	AUC.↑	Acc.↑	F1.↑	CE.↓	AUC.↑
LLM-Based	Zeroshot	gemini-2.0-flash	0.5920	0.2940	0.9257	0.6561	0.5800	0.5092	0.8328	0.7607
		GPT-4o-mini	0.6300	0.2757	2.7258	0.3657	0.5433	0.5387	0.8562	0.7395
	Zeroshot-CoT	gemini-2.0-flash	0.5880	0.3478	0.9415	0.6331	0.5800	0.5073	0.8436	0.7526
		GPT-4o-mini	0.6420	0.2960	0.8957	0.6237	0.5500	0.5353	0.8540	0.7465
	Fewshot	gemini-2.0-flash	0.7580	0.7027	8.7244	0.7956	0.5667	0.5740	12.0324	0.7053
		GPT-4o-mini	0.6815	0.4945	7.0029	0.7395	0.5067	0.5097	6.6110	0.6891
	TextGrad	gemini-2.0-flash	0.5568	0.2980	1.2011	0.5400	0.4241	0.4014	5.7813	0.6363
		GPT-4o-mini	0.6500	0.3620	0.9079	0.5364	0.5084	0.4962	4.5412	0.6709
	ATHENA	gemini-2.0-flash	0.7679	0.7222	0.9041	0.8387	0.6797	0.5968	0.7610	0.8370
		GPT-4o-mini	**0.8134**	**0.7655**	1.0863	**0.8825**	**0.7345**	**0.7161**	0.7551	**0.8704**
Utility Theory	MNL	/	0.6101	0.3887	0.8353	0.7074	0.4150	0.1955	1.0510	0.4301
	CLogit	/	0.5714	0.2424	0.8916	0.5976	0.4150	0.1955	1.0510	0.5000
	Latent Class MNL	/	0.6101	0.3967	0.8175	0.7182	0.1950	0.1088	1.0986	0.5000
Machine Learning	Logistic Regression	/	0.5620	0.5570	0.9310	0.7460	0.6500	0.6690	0.7630	0.8330
	Random Forest	/	0.7100	0.7050	0.7380	0.8810	0.6300	0.6470	**0.7290**	0.8420
	XGBoost	/	0.7080	0.7050	0.7040	0.8810	0.6300	0.6480	1.1420	0.8150
	BERT	/	0.7246	0.4994	**0.7037**	0.8811	0.6350	0.6541	0.7409	0.8168
	TabNet	/	0.6375	0.4060	0.7887	0.8810	0.6650	0.6684	0.8968	0.8147
	MLP	/	0.6475	0.6386	0.7626	0.8350	0.6068	0.6062	0.9320	0.8205

uncertain nature of human decision-making, which our model is designed to accommodate. Overall, the performance enhancements highlight ATHENA's strength in combining symbolic structures with semantic adaptation for effective personalized decision modeling.

Disentangling Semantically Similar Choices. Prompt-only LLMs and classical choice models frequently fail to distinguish between superficially similar options. For example, the few-shot LLM misclassified 75% of true *Car* trips as the premium *Metro* service. By introducing symbolic-level structure and performing individual-level semantic adaptation, ATHENA more than doubled the number of correctly classified *Car* trips, while maintaining high recall for both *Train* and *Metro*. On the Vaccine task, its learned templates encode key interactions such as age-risk trade-offs and prior-infection hesitancy, allowing it to achieve the highest F1 score despite strong semantic similarity between fully vaccinated and booster options. In practice, these interpretable templates enable a better understanding of individual behavior, for instance, identifying who tends to decline vaccination and why, which is crucial for informing high-stakes decision-making. See Appendix A.3 for details.

Computational Complexity and Scalability. With T and T' fixed, ATHENA's runtime is linear in the number of groups $|\mathcal{G}|$ and individuals N, scaled by the average LLM latency τ:

$$\mathcal{O}((|\mathcal{G}|KT + NT')\,\tau_{\text{tok}}).$$

Both stages parallelize naturally, as group-level searches run independently and individual-level refinements can be batched or distributed. Detailed runtime measurements are provided in Appendix D.

Extended backbone LLM comparisons. On a 100-individual subset, we also tested larger reasoning LLMs (`Qwen3-32B`, `DeepSeek-R1-Distill-Qwen-32B`, `GPT-4o`). With prompt-only baselines, larger reasoning models occasionally yield higher F1/Acc but exhibit volatile calibration (high CE), reflecting the lack of structural constraints. Under ATHENA, backbone differences shrink: the symbolic discovery plus semantic adaptation turns the task into constrained sampling and small, directed improvements, allowing lightweight models to reach near-maximal performance, while stronger reasoning models provide modest, consistent gains on harder interactions (e.g., vaccine risk–trust trade-offs). Full experimental details appear in Appendix C for completeness.

4.3 Ablation Study

We evaluate ATHENA's two components by *(i)* keeping only the group-level symbolic utility discovery and *(ii)* keeping only the individual-level semantic adaptation, under identical data and metrics. We do not include a symbolic-only group-level discovery baseline (Stage 1 without LLM), because the Concept Library is accessible only via the LLM. Excluding it would reduce the hypothesis space to symbolic operators alone, changing the problem definition rather than providing a clean ablation.

Group-Level Symbolic Utility Discovery: necessary but not sufficient. When ATHENA retains only the group-level symbolic component, accuracy exceeds the classical MNL by 4.7% on *Swissmetro*

and 19% on *Vaccine* (Table 2), indicating that only LLM-generated utility expressions can already encode broad demographic regularities. The accuracy trajectories of this symbolic discovery process over 30 iterations (Fig. 3) further demonstrate its effectiveness, illustrating the gradual learning of these group-level trends. Nevertheless, lower F1 score and AUC and elevated cross-entropy, reflecting limited discriminative capacity for similar alternatives. These results highlight the symbolic stage's strength in pruning the hypothesis space to interpretable structures, but also expose its limitations in capturing much heterogeneity.

Table 2: Component-wise ablation results on the Swissmetro and Vaccine choice tasks, comparing Symbolic Utility Discovery only, Semantic Adaptation only, MNL, and the full ATHENA pipeline.

		Swissmetro				Vaccine			
	Variant	**Acc.↑**	**F1.↑**	**CE.↓**	**AUC.↑**	**Acc.↑**	**F1.↑**	**CE.↓**	**AUC.↑**
Ablation Variants	Symbolic Utility Discovery Only	0.6566	0.3785	2.6044	0.5687	0.6067	0.3596	1.0410	0.7294
	Semantic Adaptation Only	0.6044	0.4950	2.2897	0.6872	0.5433	0.5348	0.8695	0.7535
	MNL	0.6101	0.3967	0.8175	0.7182	0.4150	0.1955	1.0510	0.5000
	Full Pipeline	**0.8134**	**0.7655**	**1.0863**	**0.8825**	**0.7345**	**0.7161**	**0.7551**	**0.8704**

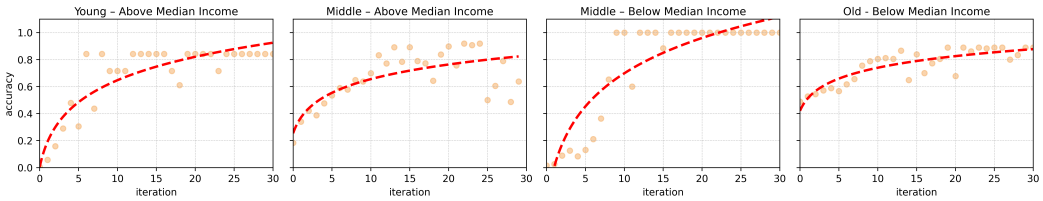


Figure 3: **Accuracy trajectories of symbolic regression.** As shown here, the accuracy keeps growing in 30 iterations for all four groups in the *vaccine* dataset. Each orange dot is the average accuracy at a given iteration; the red dashed curve is a fit showing the overall upward trend and convergence.

Individual-Level Semantic Adaptation: powerful only with a solid starting point. Conversely, bypassing symbolic discovery and initiating TextGrad from random templates leads to noteworthy degraded performance: As shown in Table 2, accuracy drops to 60.4% on the Swissmetro dataset and 54.3% on the Vaccine dataset; Swissmetro's CE more than doubles (2.29), and AUC falls below 0.70. Without a sound starting point, gradients are likely to converge to local optima and yield erratic probability outputs, reaffirming the unreliability of unguided adaptation in multi-choice settings.

Take-away. The two stages of ATHENA are complementary: symbolic discovery supplies an interpretable, well-regularized search space, while semantic adaptation injects the individual-level nuance that symbolic rules alone miss.

4.4 Symbolic Utility Discovery Fragment Analysis

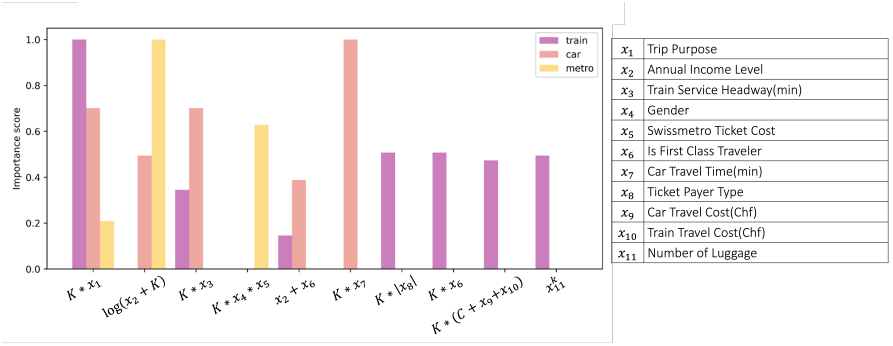
Equation (8) shows that an individual prediction is influenced by group-level symbolic utility $f_g^*(X_i; \theta_g^*)$. In this section, we demonstrate the building blocks of those utilities are both behaviorally meaningful and reusable across groups. As shown in Fig. 4, each symbolic utility is decomposed into atomic fragments $\{\varphi_1, \varphi_2, \dots\}$ and their global importance is quantified.

Fragment score. For every group g we retain the top-K ($K = 3$) utilities ranked by held-out accuracy $\mathrm{Acc}(f)$. The importance score of a fragment φ_m is then

$$\mathrm{Score}\,(\varphi_m) = \sum_{g \in \mathcal{G}} \sum_{k=1}^{K} \mathbb{1}\left[\varphi_m \subset \left\{f_{g,k}^*\right\}\right] \cdot \mathrm{Acc}\left(f_{g,k}^*\right) \tag{9}$$

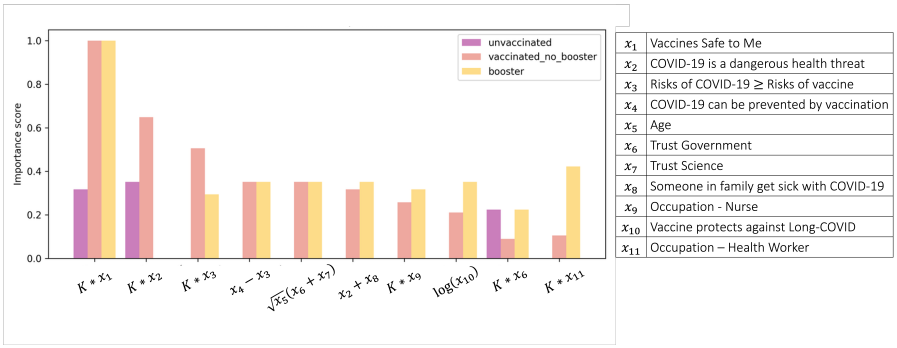
So a fragment earns points whenever it (i) appears in the top-ranked utilities of *many* groups and (ii) is embedded in highly predictive expressions.

Fig. 4 visualizes the fragment scores for both datasets. Only a small fraction of fragments dominate, confirming that ATHENA converges to a compact and interpretable symbolic basis. For example, in *Vaccine*, one of the leading fragment $\varphi_7 = \sqrt{\mathrm{Age}} * (\mathrm{Trust\ Government} + \mathrm{Trust\_Science})$ softens



x_1	Trip Purpose
x_2	Annual Income Level
x_3	Train Service Headway(min)
x_4	Gender
x_5	Swissmetro Ticket Cost
x_6	Is First Class Traveler
x_7	Car Travel Time(min)
x_8	Ticket Payer Type
x_9	Car Travel Cost(Chf)
x_{10}	Train Travel Cost(Chf)
x_{11}	Number of Luggage

(a) Swissmetro travel-mode choice dataset



x_1	Vaccines Safe to Me
x_2	COVID-19 is a dangerous health threat
x_3	Risks of COVID-19 ≥ Risks of vaccine
x_4	COVID-19 can be prevented by vaccination
x_5	Age
x_6	Trust Government
x_7	Trust Science
x_8	Someone in family get sick with COVID-19
x_9	Occupation - Nurse
x_{10}	Vaccine protects against Long-COVID
x_{11}	Occupation – Health Worker

(b) COVID-19 vaccine-choice dataset

Figure 4: **Aggregated fragment importance extracted from the learned symbolic utilities.** For each task, we plot the top-ranked atomic fragments φ_m that appear in the three best group-level utility formulas and weight them by fragment importance score 9. Values shown here are the normalized scores in $[0, 1]$.

age's impact at higher values while amplifying it for individuals who trust government or science, precisely isolating the cohort most likely to take boosters.

Beyond fragment-level analysis, ATHENA also produces fully interpretable symbolic utilities. Representative full formulas and domain-relevant insights for both *Swissmetro* and *Vaccine* are provided in Appendix A.4.

5 Conclusion

This research highlights the critical role of textual-semantic information in overcoming the limitations of traditional utility-based models for human decision-making. By introducing ATHENA, an adaptive textual-symbolic and human-centric reasoning framework is proposed that integrates group-level symbolic regression of utility functions with individual-level, LLM-powered semantic modeling, we offer a more comprehensive and personalized view of choice behavior. Our experiments on transportation mode choice and vaccine uptake demonstrate that this co-design approach clearly outperforms three existing model zoos, including classical utility, machine learning, and purely LLM-based approach, underscoring the benefits of capturing both structured attributes and rich semantic context. These findings suggest that textualized-symbolic reasoning can bridge the gap between theoretical utility optimization and real-world individual choices, paving the way for more adaptive and human-centric decision models.

Limitations. The current implementation of ATHENA has two limitations. 1) Computational Complexity: The proposed framework requires extra computational resources for textual gradient, particularly when scaling to larger populations. 2) Representation on Groups: ATHENA assumes that a shared symbolic utility function can effectively model each demographic group. However, groups with greater internal diversity may produce weaker or less reliable representations. 3) Result Stability: All reported results are based on single representative runs under fixed random seeds, given the computational cost of multi-stage adaptation. Future work will include multi-seed repetitions to further examine the stability of ATHENA's performance.

Acknowledgments

Yang Zhao acknowledges a fellowship from JHU + Amazon Initiative for Interactive AI.

References

[1] Feng Fu, Daniel I Rosenbloom, Long Wang, and Martin A Nowak. Imitation dynamics of vaccination behaviour on social networks. *Proceedings of the Royal Society B: Biological Sciences*, 278(1702):42–49, 2011.

[2] Lyndal Bond and Terry Nolan. Making sense of perceptions of risk of diseases and vaccinations: a qualitative study combining models of health beliefs, decision-making and risk perception. *BMC public health*, 11:1–14, 2011.

[3] Cornelia Betsch, Robert Böhm, Lars Korn, and Cindy Holtmann. On the benefits of explaining herd immunity in vaccine advocacy. *Nature human behaviour*, 1(3):0056, 2017.

[4] John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior: 60th anniversary commemorative edition. In *Theory of games and economic behavior*. Princeton university press, 2007.

[5] Matihew Rabin. Risk aversion and expected-utility theory: A calibration theorem. In *Handbook of the fundamentals of financial decision making: Part I*, pages 241–252. World Scientific, 2013.

[6] John Quiggin. *Generalized expected utility theory: The rank-dependent model*. Springer Science & Business Media, 2012.

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

[8] Kenneth A Small and Harvey S Rosen. Applied welfare economics with discrete choice models. *Econometrica: Journal of the Econometric Society*, pages 105–130, 1981.

[9] Milad Haghani, Michiel CJ Bliemer, and David A Hensher. The landscape of econometric discrete choice modelling research. *Journal of choice modelling*, 40:100303, 2021.

[10] Gerardo Berbeglia, Agustín Garassino, and Gustavo Vulcano. A comparative empirical study of discrete choice models in retail operations. *Management Science*, 68(6):4005–4023, 2022.

[11] Yunhan Zheng, Shenhao Wang, and Jinhua Zhao. Equality of opportunity in travel behavior prediction with deep neural networks and discrete choice models. *Transportation Research Part C: Emerging Technologies*, 132:103410, 2021.

[12] Patricio Salas, Rodrigo De la Fuente, Sebastian Astroza, and Juan Antonio Carrasco. A systematic comparative evaluation of machine learning classifiers and discrete choice models for travel mode choice in the presence of response heterogeneity. *Expert Systems with Applications*, 193:116253, 2022.

[13] Mostafa Ameli, Mohamad Sadegh Shirani Faradonbeh, Jean-Patrick Lebacque, Hossein Abouee-Mehrizi, and Ludovic Leclercq. Departure time choice models in urban transportation systems based on mean field games. *Transportation Science*, 56(6):1483–1504, 2022.

[14] Vicki M Bier, Yuqun Zhou, and Hongru Du. Game-theoretic modeling of pre-disaster relocation. *The Engineering Economist*, 65(2):89–113, 2020.

[15] Lixu Li, Zhiqiang Wang, and Xiaoqing Xie. From government to market? a discrete choice analysis of policy instruments for electric vehicle adoption. *Transportation Research Part A: Policy and Practice*, 160:143–159, 2022.

[16] Luis Enrique Loría-Rebolledo, Michael Abbott, Mélanie Antunes, Patricia Norwood, Mandy Ryan, Verity Watson, and Hangjian Wu. Public preferences and willingness to pay for a net zero nhs: a protocol for a discrete choice experiment in england and scotland. *BMJ open*, 14 (6):e082863, 2024.

[17] Tobias Thomas, Dominik Straub, Fabian Tatai, Megan Shene, Tümer Tosik, Kristian Kersting, and Constantin A Rothkopf. Modelling dataset bias in machine-learned theories of economic decision-making. *Nature Human Behaviour*, 8(4):679–691, 2024.

[18] Paul W Glimcher. Efficiently irrational: deciphering the riddle of human choice. *Trends in cognitive sciences*, 26(8):669–687, 2022.

[19] Paul JH Schoemaker. The expected utility model: Its variants, purposes, evidence and limitations. *Journal of economic literature*, pages 529–563, 1982.

[20] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[21] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

[22] Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan Reddy. Llm-sr: Scientific equation discovery via programming with large language models, 04 2024.

[23] Arya Grayeli, Atharva Sehgal, Omar Costilla Reyes, Miles Cranmer, and Swarat Chaudhuri. Symbolic regression with a learned concept library. *Advances in Neural Information Processing Systems*, 37:44678–44709, 2024.

[24] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR, 2023.

[25] Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268, 2022.

[26] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838, 2023.

[27] Giselle Moraes Ramos, Winnie Daamen, and Serge Hoogendoorn and. A state-of-the-art review: Developments in utility theory, prospect theory and regret theory to investigate travellers' behaviour in situations involving travel time uncertainty. *Transport Reviews*, 34 (1):46–67, 2014. doi: 10.1080/01441647.2013.856356. URL `https://doi.org/10.1080/01441647.2013.856356`.

[28] Wissam Qassim Al-Salih and Domokos Esztergár-Kiss. Linking mode choice with travel behavior by using logit model based on utility function. *Sustainability*, 13(8), 2021. ISSN 2071-1050. doi: 10.3390/su13084332. URL `https://www.mdpi.com/2071-1050/13/8/4332`.

[29] Jonas De Vos, Patricia L. Mokhtarian, Tim Schwanen, Veronique Van Acker, and Frank Witlox. Travel mode choice and travel satisfaction: bridging the gap between decision utility and experienced utility. *Transportation*, 43(5):771–796, 2016. ISSN 1572-9435. doi: 10.1007/s11116-015-9619-9. URL `https://doi.org/10.1007/s11116-015-9619-9`.

[30] Dick Ettema, Margareta Friman, Tommy Gärling, and Lars E. Olsson. *Travel Mode Use, Travel Mode Shift and Subjective Well-Being: Overview of Theories, Empirical Findings and Policy Implications*, pages 129–150. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016. ISBN 978-3-662-48184-4. doi: 10.1007/978-3-662-48184-4_7. URL `https://doi.org/10.1007/978-3-662-48184-4_7`.

[31] Akshay Vij and Joan L Walker. Hybrid choice models: The identification problem. In *Handbook of choice modelling*, pages 522–567. Edward Elgar Publishing, 2024.

[32] Emily Lancsar, Denzil G Fiebig, and Arne Risa Hole. Discrete choice experiments: a guide to model specification, estimation and software. *Pharmacoeconomics*, 35:697–716, 2017.

[33] Zexuan Wang. Rational inattention: The interplay of stakes and prior beliefs in a laboratory study. *Available at SSRN 5140893*, 2025.

[34] Lijun Yu and Bin Sun. Four types of typical discrete choice models: Which are you using? In *Proceedings of 2012 IEEE International Conference on Service Operations and Logistics, and Informatics*, pages 298–301. IEEE, 2012.

[35] Khandker Nurul Habib. Rational inattention in discrete choice models: Estimable specifications of ri-multinomial logit (ri-mnl) and ri-nested logit (ri-nl) models. *Transportation Research Part B: Methodological*, 172:53–70, 2023.

[36] Yeshitila Deneke, Robel Desta, Anteneh Afework, János Tóth, et al. Transportation mode choice behavior with multinomial logit model: work and school trips. *Transactions on transport sciences*, 15(1):17–27, 2024.

[37] Saeed Rahmani, Asiye Baghbani, Nizar Bouguila, and Zachary Patterson. Graph neural networks for intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 24(8):8846–8885, 2023. doi: 10.1109/TITS.2023.3257759.

[38] Hui Zhang, Li Zhang, Yanjun Liu, and Lele Zhang. Understanding travel mode choice behavior: Influencing factors analysis and prediction with machine learning method. *Sustainability*, 15 (14):11414, 2023.

[39] Juan Pineda-Jaramillo and Óscar Arbeláez-Arenas. Assessing the performance of gradient-boosting models for predicting the travel mode choice using household survey data. *Journal of Urban Planning and Development*, 148(2):04022007, 2022.

[40] Eui-Jin Kim. Analysis of travel mode choice in seoul using an interpretable machine learning approach. *Journal of Advanced Transportation*, 2021(1):6685004, 2021.

[41] José Ángel Martín-Baos, Julio Alberto López-Gómez, Luis Rodriguez-Benitez, Tim Hillel, and Ricardo García-Ródenas. A prediction and behavioural analysis of machine learning methods for modelling travel mode choice. *Transportation research part C: emerging technologies*, 156:104318, 2023.

[42] Hui Zhang, Li Zhang, Yanjun Liu, and Lele Zhang. Understanding travel mode choice behavior: Influencing factors analysis and prediction with machine learning method. *Sustainability*, 15 (14), 2023. ISSN 2071-1050. doi: 10.3390/su151411414. URL `https://www.mdpi.com/2071-1050/15/14/11414`.

[43] Li Tang, Chuanli Tang, Qi Fu, and Changxi Ma. Predicting travel mode choice with a robust neural network and shapley additive explanations analysis. *IET Intelligent Transport Systems*, 18(7):1339–1354, 2024.

[44] Anil Koushik, M Manoj, and N Nezamuddin. Explaining deep learning-based activity schedule models using shapley additive explanations. *Transportation Letters*, 17(3):442–457, 2025.

[45] Aikun Xu, Ping Zhong, Yilin Kang, Jiongqiang Duan, Anning Wang, Mingming Lu, and Chuan Shi. Than: Multimodal transportation recommendation with heterogeneous graph attention networks. *IEEE Transactions on Intelligent Transportation Systems*, 24(2):1533–1543, 2023. doi: 10.1109/TITS.2022.3221370.

[46] Halil Çevik, Ondřej Přibyl, and Shoaib Samandar. Understanding travel behavior: A deep neural network and shap approach to mode choice determinants. *Neural Network World*, 34: 219–241, 01 2024. doi: 10.14311/NNW.2024.34.012.

[47] Zahra Sadeghi, Roohallah Alizadehsani, Mehmet Akif CIFCI, Samina Kausar, Rizwan Rehman, Priyakshi Mahanta, Pranjal Kumar Bora, Ammar Almasri, Rami S. Alkhawaldeh, Sadiq Hussain, Bilal Alatas, Afshin Shoeibi, Hossein Moosaei, Milan Hladík, Saeid Nahavandi, and Panos M. Pardalos. A review of explainable artificial intelligence in healthcare. *Computers and Electrical Engineering*, 118:109370, 2024. ISSN 0045-7906. doi: https://doi.org/10.1016/j.compeleceng.2024.109370. URL `https://www.sciencedirect.com/science/article/pii/S0045790624002982`.

[48] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1), 2021. ISSN 1099-4300. doi: 10.3390/e23010018. URL https://www.mdpi.com/1099-4300/23/1/18.

[49] Zhongang Qi, Saeed Khorram, and Fuxin Li. Visualizing deep networks by optimizing with integrated gradients, 05 2019.

[50] Chase Walker, Kenny Chen, and Rickard Ewetz. Integrated decision gradients: Compute your attributions where the model makes its decision. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:5289–5297, 03 2024. doi: 10.1609/aaai.v38i6.28336.

[51] Douglas Adriano Augusto and Helio JC Barbosa. Symbolic regression via genetic programming. In *Proceedings. Vol. 1. Sixth Brazilian symposium on neural networks*, pages 173–178. IEEE, 2000.

[52] Jinghui Zhong, Liang Feng, Wentong Cai, and Yew-Soon Ong. Multifactorial genetic programming for symbolic regression problems. *IEEE transactions on systems, man, and cybernetics: systems*, 50(11):4492–4505, 2018.

[53] Lenka Skanderova. Self-organizing migrating algorithm: review, improvements and comparison. *Artificial Intelligence Review*, 56(1):101–172, 2023.

[54] Miles Cranmer. Interpretable machine learning for science with pysr and symbolicregression.jl, 05 2023.

[55] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. Regularized evolution for image classifier architecture search. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33014780. URL https://doi.org/10.1609/aaai.v33i01.33014780.

[56] Nour Makke and Sanjay Chawla. Interpretable scientific discovery with symbolic regression: a review. *Artif. Intell. Rev.*, 57(1), January 2024. ISSN 0269-2821. doi: 10.1007/s10462-023-10622-0. URL https://doi.org/10.1007/s10462-023-10622-0.

[57] Haoran Ye, Jiarui Wang, Zhiguang Cao, Federico Berto, Chuanbo Hua, Haeyeon Kim, Jinkyoo Park, and Guojie Song. Reevo: Large language models as hyper-heuristics with reflective evolution. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=483IPG0HWL.

[58] Arya Grayeli, Atharva Sehgal, Omar Costilla-Reyes, Miles Cranmer, and Swarat Chaudhuri. Symbolic regression with a learned concept library. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 44678–44709. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4ec3ddc465c6d650c9c419fb91f1c00a-Paper-Conference.pdf.

[59] Utkarsh Mall, Cheng Perng Phoo, Mia Chiquier, Bharath Hariharan, Kavita Bala, and Carl Vondrick. Disciple: Learning interpretable programs for scientific visual discovery. 2025.

[60] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[61] 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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu,

Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran

Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[62] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

[63] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

[64] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin,

Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh

Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal,

Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov,

Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

[65] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

[66] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

[67] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods

for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR, 2023.

[68] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[69] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[70] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[71] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*, 2023.

[72] Yu Xia, Rui Wang, Xu Liu, Mingyan Li, Tong Yu, Xiang Chen, Julian McAuley, and Shuai Li. Beyond chain-of-thought: A survey of chain-of-x paradigms for llms. *arXiv preprint arXiv:2404.15676*, 2024.

[73] Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 333–356. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/00d80722b756de0166523a87805dd00f-Paper-Conference.pdf.

[74] Qiguang Chen, Libo Qin, Jiaqi Wang, Jingxuan Zhou, and Wanxiang Che. Unlocking the capabilities of thought: A reasoning boundary framework to quantify and optimize chain-of-thought. *Advances in Neural Information Processing Systems*, 37:54872–54904, 2024.

[75] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

[76] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023. URL https://arxiv.org/abs/2305.10601.

[77] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

[78] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

[79] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1, 2023.

[80] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[81] Zhen Bi, Ningyu Zhang, Yinuo Jiang, Shumin Deng, Guozhou Zheng, and Huajun Chen. When do program-of-thought works for reasoning? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17691–17699, 2024.

[82] Tianming Liu, Manzi Li, and Yafeng Yin. Can large language models capture human travel behavior? evidence and insights on mode choice. *arXiv preprint arXiv:2408.12345*, 2024.

[83] Baichuan Mo, Hanyong Xu, Dingyi Zhuang, Ruoyun Ma, Xiaotong Guo, and Jinhua Zhao. Large language models for travel behavior prediction. *arXiv preprint arXiv:2312.00819*, 2023.

[84] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, and Franck Dernoncourt. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. *arXiv preprint arXiv:2402.01981*, 2024.

[85] Maddalena Coletta, Nolan Zaslavsky, and et al. Personas improve behavioral alignment of large language models. *arXiv preprint arXiv:2403.11568*, 2024.

[86] Ollie Liu, Deqing Fu, Dani Yogatama, and Willie Neiswanger. DeLLMa: Decision making under uncertainty with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=Acvo2RGSCy`.

[87] Yiming Lu, Yebowen Hu, Hassan Foroosh, Wei Jin, and Fei Liu. STRUX: An LLM for decision-making with structured explanations. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 131–141, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-190-2. doi: 10.18653/v1/2025.naacl-short.11. URL `https://aclanthology.org/2025.naacl-short.11/`.

[88] Beibin Li, Konstantina Mellou, Bo Zhang, Jeevan Pathuri, and Ishai Menache. Large language models for supply chain optimization. *CoRR*, 2023.

[89] Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving. In *First Conference on Language Modeling*, 2024. URL `https://openreview.net/forum?id=UPE6WYE8vg`.

[90] Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Knödler, Ulrich Keller, Dieter Beule, et al. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*, 6(11): e2343689–e2343689, 2023.

[91] Charles F Manski. The structure of random utility models. *Theory and decision*, 8(3):229, 1977.

[92] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. 1972.

[93] William N Evans and W Kip Viscusi. Estimation of state-dependent utility functions using survey data. *The Review of Economics and Statistics*, pages 94–104, 1991.

[94] Thomas Bäck and Hans-Paul Schwefel. An overview of evolutionary algorithms for parameter optimization. *Evolutionary computation*, 1(1):1–23, 1993.

[95] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[96] 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.

[97] Michel Bierlaire, Kay Axhausen, and Georg Abay. The acceptance of modal innovation: The case of swissmetro. 01 2001.

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

[99] 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.

[100] Amir Ghorbani, Neema Nassir, Patricia Sauri Lavieri, Prithvi Bhat Beeramoole, and Alexander Paz. Enhanced utility estimation algorithm for discrete choice models in travel demand forecasting. *Transportation*, pages 1–28, 2025.

[101] Shadi Haj-Yahia, Omar Mansour, and Tomer Toledo. Incorporating domain knowledge in deep neural networks for discrete choice models. *Transportation Research Part C: Emerging Technologies*, 171:105014, 2025.

[102] 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.

[103] Rachel Croson and Uri Gneezy. Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–74, June 2009. doi: 10.1257/jel.47.2.448. URL `https://www.aeaweb.org/articles?id=10.1257/jel.47.2.448`.

[104] Agnieszka Tymula, Lior A. Rosenberg Belmaker, Amy K. Roy, Lital Ruderman, Kirk Manson, Paul W. Glimcher, and Ifat Levy. Adolescents' risk-taking behavior is driven by tolerance to ambiguity. *Proceedings of the National Academy of Sciences*, 109(42):17135–17140, 2012. doi: 10.1073/pnas.1207144109. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1207144109`.

[105] Ernst-Jan de Bruijn and Gerrit Antonides. Poverty and economic decision making: a review of scarcity theory. *Theory and Decision*, 92:5 – 37, 2021. URL `https://api.semanticscholar.org/CorpusID:233665419`.

[106] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[107] Baichuan Mo, Hanyong Xu, Dingyi Zhuang, Ruoyun Ma, Xiaotong Guo, and Jinhua Zhao. Large language models for travel behavior prediction. *arXiv preprint arXiv:2312.00819*, 2023.

[108] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[109] Tianming Liu, Manzi Li, and Yafeng Yin. Can large language models capture human travel behavior? evidence and insights on mode choice. *Evidence and Insights on Mode Choice (August 26, 2024)*, 2024.

[110] Pengpeng Huang, Lei Gong, Tian Lei, Jia Wang, Cheng Zhu, and Zheng Zhang. A comparative study of mnl and machine learning methods for travel mode choice of medical travel. In *International Conference on Traffic and Transportation Studies*, pages 305–313. Springer, 2024.

[111] Thi My Thanh Truong and Thi Cam Van Nguyen. Travel time attractiveness in motorcycle dominated cities: An investigation of university students' travel behavior. In *CIGOS 2021, Emerging Technologies and Applications for Green Infrastructure: Proceedings of the 6th International Conference on Geotechnics, Civil Engineering and Structures*, pages 1723–1731. Springer, 2021.

[112] Yikang Wu, Mehmet Yildirimoglu, and Zuduo Zheng. Evaluating electric micro-mobility related mode choice stated preferences: A latent class choice approach. *arXiv preprint arXiv:2504.01237*, 2025.

[113] Mohammad Tamim Kashifi, Arshad Jamal, Mohammad Samim Kashefi, Meshal Almoshaogeh, and Syed Masiur Rahman. Predicting the travel mode choice with interpretable machine learning techniques: A comparative study. *Travel Behaviour and Society*, 29:279–296, 2022.

[114] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL https://www.deeplearningbook.org.

[115] Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6679–6687, 2021.

[116] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[117] J.D. Shires and G.C. de Jong. An international meta-analysis of values of travel time savings. *Evaluation and Program Planning*, 32(4):315–325, 2009. ISSN 0149-7189. doi: https://doi.org/10.1016/j.evalprogplan.2009.06.010. URL https://www.sciencedirect.com/science/article/pii/S0149718909000548. Evaluating the Impact of Transport Projects: Lessons for Other Disciplines.

[118] Pedro Abrantes and Mark Wardman. Meta-analysis of uk values of travel time: An update. *Transportation Research Part A: Policy and Practice*, 45:1–17, 01 2011. doi: 10.1016/j.tra.2010.08.003.

[119] Yu-Chun Chang. Factors affecting airport access mode choice for elderly air passengers. *Transportation Research Part E: Logistics and Transportation Review*, 57(C):105–112, None 2013. doi: 10.1016/j.tre.2013.01.010. URL https://ideas.repec.org/a/eee/transe/v57y2013icp105-112.html.

[120] Claude Weis, Kay W. Axhausen, Robert Schlich, and René Zbinden. Models of mode choice and mobility tool ownership beyond 2008 fuel prices. *Transportation Research Record*, 2157 (1):86–94, 2010. doi: 10.3141/2157-11.

[121] Lawrence Frank, Mark Bradley, Sarah Kavage, James Chapman, and T. Lawton. Urban form, travel time, and cost relationships with tour complexity and mode choice. *Transportation*, 35: 37–54, 11 2007. doi: 10.1007/s11116-007-9136-6.

[122] Yuan Liao, Jorge Gil, Rafael Pereira, Sonia Yeh, and Vilhelm Verendel. Disparities in travel times between car and transit: Spatiotemporal patterns in cities. *Scientific Reports*, 10, 03 2020. doi: 10.1038/s41598-020-61077-0.

[123] Jorge Marques, Sofia Gomes, Mónica Ferreira, Marina Rebuá, and Hugo Marques. Generation z and travel motivations: The impact of age, gender, and residence. *Tourism and Hospitality*, 6(2), 2025. ISSN 2673-5768. doi: 10.3390/tourhosp6020082. URL https://www.mdpi.com/2673-5768/6/2/82.

[124] Manfred Green. Rational and irrational vaccine hesitancy. *Israel Journal of Health Policy Research*, 12, 03 2023. doi: 10.1186/s13584-023-00560-1.

[125] Yunha Noh, Ju Hwan Kim, Dongwon Yoon, Young June Choe, Seung-Ah Choe, Jaehun Jung, Sang-Won Lee, and Ju-Young Shin. Predictors of covid-19 booster vaccine hesitancy among fully vaccinated adults in korea: a nationwide cross-sectional survey. *Epidemiology and Health*, 44, 2022. URL https://api.semanticscholar.org/CorpusID:251255072.

[126] Nirbachita Biswas, Toheeb Mustapha, Jagdish Khubchandani, and James H. Price. The nature and extent of covid-19 vaccination hesitancy in healthcare workers. *Journal of Community Health*, 46(6):1244–1251, December 2021. doi: 10.1007/s10900-021-00984-3. Epub 2021 Apr 20.

[127] Mallory Trent, Holly Seale, Abrar Ahmad Chughtai, Daniel Salmon, and C. Raina MacIntyre. Trust in government, intention to vaccinate and covid-19 vaccine hesitancy: A comparative survey of five large cities in the united states, united kingdom, and australia. *Vaccine*, 40(17):2498–2505, 2022. ISSN 0264-410X. doi: https://doi.org/10.1016/j.vaccine.2021.06.048. URL https://www.sciencedirect.com/science/article/pii/S0264410X21007982. Pandemic Simulation, Pacific Eclipse.

[128] Jason Glanz, Nicole Wagner, Komal Narwaney, Courtney Kraus, Jo Shoup, Stanley Xu, Sean O'Leary, Saad Omer, Kathy Gleason, and Matthew Daley. Web-based social media intervention to increase vaccine acceptance: A randomized controlled trial. *Pediatrics*, 140:e20171117, 11 2017. doi: 10.1542/peds.2017-1117.

[129] Sophie Lohmann and Dolores Albarracin. Trust in the public health system as a source of information on vaccination matters most when environments are supportive. *Vaccine*, 40, 06 2022. doi: 10.1016/j.vaccine.2022.06.012.

[130] Nathalie Bajos, Alexis Spire, Léna Silberzan, Antoine Sireyjol, Florence Jusot, Laurence Meyer, Jeanna-Eve Pousson, and Josiane Warszawski. When lack of trust in the government and in scientists reinforces social inequalities in vaccination against covid-19. *Frontiers in Public Health*, 10:908152, 07 2022. doi: 10.3389/fpubh.2022.908152.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please refer to Section 3.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Yes, we have full proof and assumption for each result. Please refer to Section 3,4.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Yes, all the details are illustrated in Section 3 and Appendix. We also have released all the code and experiment raw results on GitHub. While the experiments are reproducible given the released code and data, all reported results are from single runs. Multi-seed repetitions will be conducted in future work to assess the stability of the reported performance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Yes, the code and data will be published on GitHub upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars due to the high cost of repeated API calls. We plan to incorporate more thorough uncertainty quantification in future work as computational resources allow.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, please refer to Section 1 and Section 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets used are correctly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release the full implementation of our framework. All assets will be made publicly available upon publication.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The study uses publicly available datasets involving human subjects. Based on the dataset documentation, the original data collection was conducted under appropriate ethical review and consent procedures.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are used in both stages of our method: to generate candidate symbolic utility expressions during the group-level modeling phase, and to initialize individual-level templates during semantic adaptation.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Extended Analysis

A.1 Broader Societal Impacts

The introduction of ATHENA, a two-stage framework that first discovers group-level symbolic utility functions and then develops individual semantic templates, offers potential for positive societal contributions. By combining LLM reasoning with symbolic regression, ATHENA aims to deliver: **(1) More inclusive public-policy insights:** interpretable utility functions reveal group motivations behind choices in socially critical behaviors, enabling better targeted interventions and policy making. **(2) Precise and equitable individual adaptation:** reduce the "one-size-fits-all" errors in some high-stake domains, like healthcare and education.

At the same time, ATHENA also brings concerns. (1) Potential Misuse: Because ATHENA can model fine-grained individual decisions, it could be deployed for manipulative advertising, political micro-targeting, or discriminatory dynamic pricing. Mitigations include restricted licensing and mandatory human oversight for high-impact deployments. (2) Bias and Fairness: The symbolic utility discovery stage may suffer from coarse group partitioning, which can ignore intra-group heterogeneity. Meanwhile, the semantic adaptation stage may inherit biases presented in the LLM's training data, thereby embedding historical stereotypes.

A.2 Qualitative Analysis and Case Study

Table 3: **Qualitative case study of ATHENA on the *Swissmetro* dataset.** This table contrasts four representative travelers' attributes, the candidate alternatives, the group-level symbolic utility functions learned in Stage-1, and the individualized decision rules refined in Stage-2.

Swissmetro **Case 1**	
Features $\mathcal{X}$	- **Age**: between 54 and 65 years old - **Gender**: Male - **Income**: Over 100 - **Trip Purpose**: business - **Luggage**: none of luggage - **Payment Method**: paid by oneself - **Origin**: St. Gallen - **Destination**: Bern
Alternatives $\mathcal{J}$	- **Metro**: travel time of 77 minutes, costing 74 CHF, with a headway of 30 minutes - **Train**: travel time of 120 minutes, costing 64 CHF, with a headway of 120 minutes - **Car**: travel time of 169 minutes, costing 60 CHF
Optimal Group Utility f_g^*	- **Train**: $K_1 \cdot \text{trip\_purpose} + K_2 \cdot (\sqrt{\text{num\_luggage} + C_1} + C_2) \cdot \log(\text{age} + C_3) + K_3 \cdot \text{time\_train} + K_4 \cdot \sqrt{\text{train\_headway\_min}} + K_5 \cdot (\text{is\_first\_class} + C_4) \cdot \sqrt{\log(\text{income} + C_5)} + C_6$ - **Car**: $K_1 \cdot \text{trip\_purpose} + K_2 \cdot \sqrt{\text{num\_luggage} + C_1} \cdot \log(\text{age} + C_2) + K_3 \cdot \text{time\_car} \cdot \sqrt{\log(\text{income} + C_3)} + K_4 \cdot \sqrt{\lvert\text{cost\_train} + C_4\rvert} + K_5 \cdot \text{train\_service\_headway} + K_6 \cdot (\text{is\_first\_class} + C_5) \cdot \sqrt{\log(\text{income} + C_6)} + C_7$ - **Metro**: $K_1 \cdot \text{trip\_purpose} + K_2 \cdot (\text{num\_luggage} + C_1) \cdot \sqrt{\log(\text{age} + C_2)} + K_3 \cdot \text{time\_sm} \cdot \sqrt{\text{income} + C_3} + K_4 \cdot \sqrt{\text{cost\_sm} + C_4} + K_5 \cdot (\text{ticket\_payer\_type} + C_5) + C_6$
Optimal Personalized Decision Rule $\mathcal{P}_i^*$	- **BALANCED**: As a business traveler, my travel decisions are guided by a dynamic scoring system that prioritizes speed, environmental impact, and cost. I often weigh these factors differently based on specific scenarios (e.g., major conferences vs. regular trips). Real-time traffic/weather and reflections on past trips refine preferences; I consider Swissmetro, trains, cars, buses, and rideshares, evaluating comfort, reliability, and amenities.

Swissmetro – Case 2	
Features $\mathcal{X}$	- **Age**: between 39 and 54 years old - **Gender**: Male - **Income**: between 50 and 100 - **Trip Purpose**: leisure - **Luggage**: no luggage - **Payment Method**: paid by oneself - **Origin**: Graubünden - **Destination**: Bern
Alternatives $\mathcal{J}$	- **Metro**: travel time 142 min, cost 123 CHF, headway 30 min - **Train**: travel time 180 min, cost 97 CHF, headway 60 min - **Car**: travel time 136 min, cost 149 CHF
Optimal Group Utility f_g^*	- **Train**: $K_1 \cdot (\text{ticket\_payer\_type} \cdot \text{is\_car\_available} \sqrt{\text{age} + C_1} + K_2 \cdot \text{is\_first\_class} \cdot \log(\text{age} + C_2)) - K_3 \cdot \frac{\text{time\_train}}{\text{age} + C_3} + K_4 \cdot (\text{income} + \text{num\_luggage} \sqrt{\text{age} + C_4}) - K_5 \cdot \left(\frac{\text{time\_car}}{\text{age} + C_5} + C_6 \right) + C_7$ - **Car**: $K_1 \cdot (\text{ticket\_payer\_type} \cdot \text{is\_car\_available} \sqrt{\text{age} + C_1} + K_2 \cdot \text{is\_first\_class}) - K_3 \cdot \frac{\text{time\_train}}{\text{age} + C_2} + K_4 \cdot \log(\text{income} + \text{num\_luggage} + K_5 \sqrt{\text{age} + C_3}) - K_6 \cdot \left(\frac{\text{time\_car}}{\text{age} + C_4} + C_5 \right) + C_6$ - **Metro**: $K_1 \cdot (\text{sm\_headway\_min} - \frac{\text{time\_sm}}{\log(\text{age} + C_1)}) + K_2 \cdot (\text{has\_ga\_travel\_pass} \sqrt{\text{age} + C_2} + \text{gender} \cdot \log(\text{income} + \text{num\_luggage} + C_3)) + C_4$
Optimal Personalized Decision Rule $\mathcal{P}_i^*$	- **BALANCED, INFORMED, SUSTAINABLE, USER-CENTRIC, CONTEXT-ADAPTIVE, COMFORT-FOCUSED**: Dynamically trades off time, cost, and scenic enjoyment. Eco-friendliness and comfort (leg-room, quiet cars) matter, even with longer trips or higher prices; prior-trip feedback and real-time context refine recommendations.
Swissmetro – Case 3	
Features $\mathcal{X}$	- **Age**: between 39 and 54 years old - **Gender**: Male - **Income**: under 50 - **Trip Purpose**: leisure - **Luggage**: no luggage - **Payment Method**: paid by unknown people - **Origin**: Zurich - **Destination**: Bern
Alternatives $\mathcal{J}$	- **Metro**: travel time 56 min, cost 42 CHF, headway 10 min - **Train**: travel time 111 min, cost 36 CHF, headway 30 min - **Car**: travel time 88 min, cost 60 CHF

Optimal Group Utility f_g^*	- **Train**: $C_1 + K_1 \cdot \lvert\text{time\_train} - \text{time\_car}\rvert(\log(\text{income} + C_2) + \sqrt{\text{age} + C_3})(\text{gender}\sqrt{\text{num\_luggage} + C_4} + \text{is\_first\_class}) + K_2 \cdot \frac{\text{trip\_purpose}}{\sqrt{\text{age} + C_5} + C_6} - K_3 \cdot \text{cost\_train} \cdot \text{trip\_purpose}$ - **Car**: $C_1 + K_1 \cdot (\text{num\_luggage} \cdot \text{trip\_purpose} \cdot \lvert\log(\text{income} + C_2)\rvert \cdot (C_3 + \sqrt{\text{age} + C_4}) + \sqrt{\text{age} + C_5} \cdot \lvert\text{time\_train} - \text{time\_car}\rvert \cdot (C_6 + \text{is\_first\_class})) - K_2 \cdot (C_7 + \text{train\_service\_headway\_min}) + K_3 \cdot (\text{income} + C_8)^{C_9}$ - **Metro**: $C_1 + K_1 \cdot \left(\frac{\text{time\_sm}^{C_2} \cdot \text{trip\_purpose} \cdot \sqrt{\text{num\_luggage} + C_3} \cdot \log(\text{income} + C_4)}{\sqrt{\text{age} + C_5} + C_6} + K_2 \cdot (\text{cost\_sm} \cdot (\log(\text{age} + C_7) + \text{is\_first\_class})) \cdot (C_8 + \sqrt{\text{time\_car}}) \right)$
Optimal Personalized Decision Rule $\mathcal{P}_i^*$	- **COST_SAVING**: Prefers options at least 10% cheaper than peak-fare averages, values minimal transfers and amenities (Wi-Fi, food), uses carbon-footprint data, and favors off-peak scheduling; feedback refines future recommendations.
***Swissmetro* – Case 4**	
Features $\mathcal{X}$	- **Age**: over 65 years old - **Gender**: Male - **Income**: over 100 - **Trip Purpose**: shopping - **Luggage**: no luggage - **Payment Method**: paid half-half - **Origin**: Vaud - **Destination**: Geneva
Alternatives $\mathcal{J}$	- **Metro**: travel time 21 min, cost 226 CHF, headway 10 min - **Train**: travel time 42 min, cost 209 CHF, headway 60 min - **Car**: travel time 40 min, cost 67 CHF
Optimal Group Utility f_g^*	- **Train**: $K_1 \cdot (\text{time\_train} + \text{num\_luggage} \cdot \text{age}^{C_1} + \text{age}^{C_2} + \text{time\_car}) + K_2 \cdot (\text{income} \cdot \text{is\_first\_class} \cdot \text{gender} \cdot \log(\text{age} + C_3)) + \text{num\_luggage}^{C_4} + C_5$ - **Car**: $K_1 \cdot (\text{time\_car} + \text{num\_luggage} \cdot \text{age}^{C_1} + \text{income}\sqrt{\text{is\_first\_class}} + \text{car\_travel\_cost\_chf} + \text{num\_luggage} \cdot \exp(\text{age}/C_2)) + C_3$ - **Metro**: $K_1 \cdot \text{time\_sm} + K_2 \cdot (\text{cost\_sm} + \text{income} \cdot (\text{is\_first\_class} + \text{gender}) \cdot \exp(\text{age}^{C_1}) + \text{num\_luggage}\,\text{age}^{C_2} + \text{age}/C_3) + C_4$
Optimal Personalized Decision Rule $\mathcal{P}_i^*$	- **COMFORT_SEEKING**: Prioritizes spacious seating, quiet cars, and onboard services; willing to pay up to 20% premium. Prefers real-time updates and easy boarding for accessibility; social events may nudge to more social modes; feedback refines future recommendations.

Table 4: **Qualitative case study of ATHENA on the *Vaccine* dataset.** This table contrasts four representative individuals' attributes, the candidate alternatives, the group-level symbolic utility functions learned in Stage-1, and the individualized decision rules refined in Stage-2.

***Vaccine* – Case 1**	

Features $\mathcal{X}$	- **Age**: 25 - **Gender**: Male - **Occupation**: Nurse - **Education**: No university degree - **Income**: Above-median - **COVID-19 Threat Perception**: Moderate - **Risk Perception**: Disease risk $>$ vaccine risk - **Trust in Government**: Moderate - **Trust in Science**: Moderate - **Perceived Vaccine Safety**: Fairly safe - **Family COVID Infection**: $>$1 yr ago - **Attention to Vaccine News**: Increased
Alternatives $\mathcal{J}$	- **Unvaccinated** - **Vaccinated_No_Booster** - **Booster**
Optimal Group Utility f_g^*	- **Unvaccinated**: $C_1 \cdot \text{covid\_threat}(C_2 + \text{trust\_gov} \cdot \text{trust\_sci} \cdot \log(\text{age} + C_3)) \cdot \text{risk\_covid\_gt\_vax} + K_1 \cdot \text{family\_covid} \cdot \log(\text{age} + C_4)$ - **Vaccinated_No_Booster**: $C_1 \cdot \text{covid\_threat} + C_2 \cdot \text{vax\_safe} + K_1 \cdot (\text{trust\_gov} \cdot \text{trust\_sci} \cdot \text{more\_attention} \sqrt{\text{age} + C_3})$ - **Booster**: $C_1 \cdot e^{\text{age}^{C_2}}$ covid_threat $\sqrt{\text{vax\_protect\_long}} + C_3 \cdot \text{vax\_safe} + K_1 \cdot (\text{trust\_gov} \cdot \text{trust\_sci} \cdot \text{nurse} \cdot \sqrt{\text{age} + C_4})$
Optimal Personalized Decision Rule $\mathcal{P}_i^*$	- **TRUSTING_AUTHORITY**: This persona represents a cautiously informed healthcare worker who values evidence-based guidance and may favor "Vaccinated (No Booster)" given safety concerns, while remaining open to updates as new data emerge; family and social influence are considered.
***Vaccine* – Case 2**	
Features $\mathcal{X}$	- **Age**: 55 - **Gender**: Male - **Education**: No university degree - **Income**: Below-median - **COVID-19 Threat Perception**: Moderate - **Trust in Government Delivery**: High - **Trust in Science**: Some - **Risk Perception**: Disease risk $>$ vaccine risk - **Family COVID**: None - **Attention to Vaccine News**: Decreased
Alternatives $\mathcal{J}$	- **Unvaccinated** - **Vaccinated_No_Booster** - **Booster**
Optimal Group Utility f_g^*	- **Unvaccinated**: $K_1 \sqrt{\text{covid\_threat}(\text{risk\_covid\_gt\_vax} + \text{gender} \sqrt{\text{age}} + C_1)} \cdot ((\text{trust\_gov}\, \text{trust\_sci})^{C_2} + C_3) + K_2 \cdot \text{more\_attention} - K_3 \cdot \text{low\_income}(C_4 + \text{has\_degree} \cdot \text{trust\_gov} \cdot \text{trust\_sci}) + C_5$ - **Vaccinated_No_Booster**: $K_1 \cdot (\text{vax\_safe} + \text{trust\_gov} \cdot \text{trust\_sci} \cdot \sqrt{\sqrt{\text{age} + C_1} + \text{income\_unknown} + C_2} + C_3) + K_2 \cdot \frac{\text{more\_attention}}{\text{less\_attention} + C_4} + C_5$ - **Booster**: $K_1 \cdot (\text{family\_covid} + \text{physician} \cdot \text{trust\_gov}\, \text{trust\_sci}\, (\sqrt{\text{age} + C_1} + C_2) + \text{nurse} \cdot \text{trust\_sci} \sqrt{\text{age} + C_3} + C_4) + C_5$

Optimal Personalized Decision Rule $\mathcal{P}_i^*$	- **SKEPTICAL**: Prefers conservative choices due to perceived safety concerns; may remain unvaccinated unless convinced by trusted figures; open to "Vaccinated_No_Booster" or "Booster" if necessity and safety are clearly established.
***Vaccine* – Case 3**	
Features $\mathcal{X}$	- **Age**: 86 - **Gender**: Male - **Education**: No university degree - **Income**: Below-median - **COVID-19 Threat View**: Moderate - **Perceived Vaccine Safety**: High - **Long-COVID Protection Belief**: Uncertain - **Family COVID**: None - **Attention to Vaccine News**: Increased
Alternatives $\mathcal{J}$	- **Unvaccinated** - **Vaccinated_No_Booster** - **Booster**
Optimal Group Utility f_g^*	- **Unvaccinated**: $C_1 + K_1 \cdot (\text{covid\_threat} \cdot \text{trust\_gov} \cdot \text{trust\_sci} \cdot \text{age}^{C_2}(C_3 + \text{family\_covid})) - K_2 \cdot (\text{risk\_covid\_gt\_vax}(C_4 + \text{income\_unknown} \cdot e^{-K_3 \cdot \text{more\_attention age}^{C_5}}))$ - **Vaccinated_No_Booster**: $C_1 - K_1 \cdot (\text{age}^{C_2}(C_3 + \text{low\_income})(C_4 - \text{trust\_sci})) + K_2 \cdot (\text{vax\_safe trust\_gov}\,(C_5 + \text{more\_attention age}^{C_6}))$ - **Booster**: $C_1 + K_1 \cdot (\text{vax\_protect\_long} \cdot e^{-K_2(\text{age}^{C_2} + \text{low\_income} \cdot \text{family\_covid})}) - K_3 \cdot (\text{less\_attention} \cdot \text{trust\_sci}\,(\text{age}/C_3))$
Optimal Personalized Decision Rule $\mathcal{P}_i^*$	- **THREAT_AVOIDING**: Perceives high disease risk; ranks *Booster* > *Vaccinated_No_Booster* > *Unvaccinated*; considers logistics and side-effect concerns while relying on trusted sources.
***Vaccine* – Case 4**	
Features $\mathcal{X}$	- **Age**: 52 - **Gender**: Male - **Education**: No university degree - **Income**: Below-median - **COVID-19 Threat Perception**: Strong - **Belief in Vaccine Prevention**: Low - **Trust in Science**: Moderate - **Risk Perception**: Disease risk > vaccine risk - **Attention to Vaccine News**: Unchanged - **Family COVID**: None
Alternatives $\mathcal{J}$	- **Unvaccinated** - **Vaccinated_No_Booster** - **Booster**

Optimal Group Utility f_g^*	- **Unvaccinated**: $K_1\sqrt{\text{covid\_threat}(\text{risk\_covid\_gt\_vax} + \text{gender}\sqrt{\text{age}} + C_1)} \cdot ((\text{trust\_gov}\,\text{trust\_sci})^{C_2} + C_3) + K_2 \cdot \text{more\_attention} - K_3 \cdot \text{low\_income}(C_4 + \text{has\_degree} \cdot \text{trust\_gov} \cdot \text{trust\_sci}) + C_5$ - **Vaccinated_No_Booster**: $K_1 \cdot (\text{vax\_safe} + \text{trust\_gov} \cdot \text{trust\_sci} \cdot \sqrt{\sqrt{\text{age} + C_1} + \text{income\_unknown} + C_2} + C_3) + K_2 \cdot \frac{\text{more\_attention}}{\text{less\_attention} + C_4} + C_5$ - **Booster**: $K_1 \cdot (\text{family\_covid} + \text{physician} \cdot \text{trust\_gov} \cdot \text{trust\_sci}\,(\sqrt{\text{age} + C_1} + C_2) + \text{nurse} \cdot \text{trust\_sci}\sqrt{\text{age} + C_3} + C_4) + C_5$
Optimal Personalized Decision Rule $\mathcal{P}_i^*$	- **BALANCED**: Cautious yet data-driven; moderate trust in authorities; open to boosters with clear evidence; weighs prior experiences and accessibility.

A.3 Semantically Similar Choices Analysis

As illustrated in Figure 5, ATHENA not only raises aggregated accuracy but also improves decision-critical boundaries, offering more reliable evidence for public-health and transport-policy planning.

A.4 Extended Interpretability Showcase

We provide representative full symbolic utilities discovered by ATHENA on *Swissmetro* and *Vaccine* datasets. These examples illustrate how the symbolic structure translates into actionable insights for transportation and public health domains.

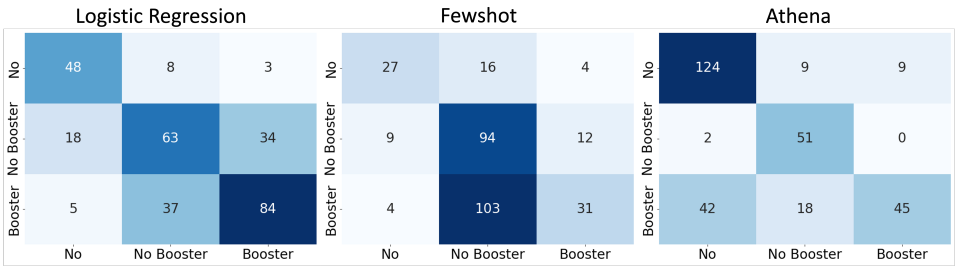
A.4.1 Representative Example — Swissmetro Dataset

Mode	Discovered symbolic utility
Train	$K_1 \cdot (\text{train\_time} + \text{metro\_time} + \text{luggage} \cdot \log(\text{age} + C_1) + \text{age} + \text{is\_male}) + C_2 \cdot (\text{first\_class} + \text{income}) - C_3 \cdot (\text{GA\_pass} + \text{headway})$
Car	$K_1 \cdot (\text{car\_time} + \text{train\_time} + \text{luggage} \cdot \log(\text{age} + C_1) + \text{age}) + C_2 \cdot (\text{first\_class} + \text{income}) - C_3 \cdot (\text{GA\_pass} + \text{metro\_fare} + \text{is\_male})$
Metro	$K_1 \cdot (\text{metro\_time} + \text{luggage} + \text{age} + \text{is\_male}) + C_2 \cdot (\text{first\_class} + \text{income}) - C_3 \cdot (\text{headway} + \text{GA\_pass} + \text{is\_male})$

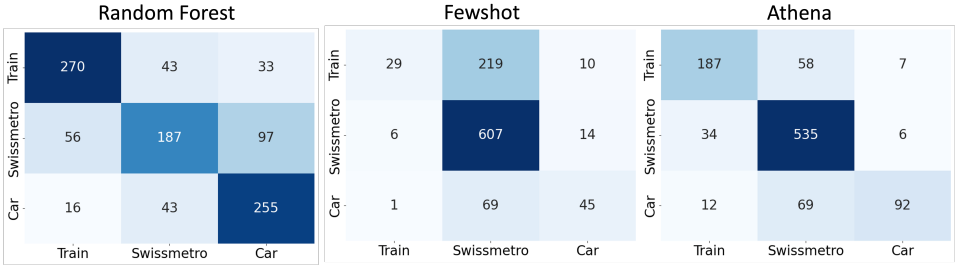
Feature: Between 39 and 54 years old, identify as female, and have an income between 50 and 100.

Key take-aways for domain experts

- **Time dominates.** Large negative coefficients on travel-time variables show this segment is **highly time-sensitive** → investments that shorten door-to-door time (e.g., skip-stop service) should shift demand [117].
- **Comfort premium.** Positive weight on $(\text{first\_class} + \text{income})$ across all modes indicates a willingness to pay for comfort that scales with income → targeted upselling (seat reservations, quiet cars) is effective [118].
- **Luggage burden grows with age.** The interaction $\text{luggage} \cdot \log(\text{age} + C_1)$ reveals baggage becomes disproportionately painful for older travelers → facilities such as luggage trolleys or porter services may raise train/metro share [119].
- **GA pass effect.** Owning a GA pass biases travellers away from modes that still incur extra fares. Extending GA coverage to Swissmetro would therefore raise its relative appeal [120].



(a) **Vaccine-uptake task.** ATHENA removes all *34* cases in which the *Vaccinated_no_booster* class was previously misclassified as *Booster*, thereby preserving the integrity of booster-demand estimates.



(b) **Travel-mode choice task.** ATHENA cuts the *Swissmetro*-versus-*Car* confusion from *83* to *6* instances, refining forecasts of low-carbon rail adoption.

Figure 5: ATHENA yields improvements on the classes that matter most yet were previously hard to distinguish.

Mode	Discovered symbolic utility
Train	$K_1 \cdot \Big(\text{purpose} + \lvert\text{payer\_type} \cdot C_1\rvert - \text{first\_class} + \lvert\text{luggage}\rvert \sqrt{\lvert\text{age} + C_2\rvert} + \lvert\text{train\_time} + C_3\rvert + \log(\text{income} + C_4)\Big) - C_5$
Car	$K_1 \cdot \Big(\lvert\text{car\_time} + C_1\rvert + \lvert\text{car\_time} - \text{train\_time} + C_2\rvert - \lvert\text{car\_cost} + \text{train\_cost} + C_3\rvert + \lvert\text{headway}\rvert \sqrt{\text{income} + C_4}\Big) + C_5$
Metro	$K_1 \cdot \Big(\lvert\text{metro\_time} + C_1\rvert + \lvert\text{metro\_cost} + C_2\rvert + \sqrt{\lvert\text{age} + C_3\rvert} + \log\Big(\exp(\text{income} + C_4) + C_5\Big)\Big) - C_6$

Feature: Male travelers younger than 24 years, annual income 50–100k.

Key take-aways for domain experts

- **Time still trumps money.** Travel time appears in all utilities, while fare only in Car/Metro. For under-25 travelers, each minute lost matters more than an extra franc → prioritizing faster transfers or signal priority is especially effective [121].
- **Headway frustration fuels car use.** The term headway $\cdot \sqrt{\text{income}}$ shows that infrequent trains push young people toward cars, and irritation rises with income → higher-frequency rail services can curb car switching [122].
- **First-class indifference.** The negative first_class coefficient suggests little interest in upgrades → amenities in standard class (Wi-Fi, gaming lounges) may be more persuasive than premium seating [123].

A.4.2 Representative Example — Vaccine Dataset

Mode	Discovered symbolic utility
Unvaccinated	$C_1 \cdot \text{covid\_threat} \cdot \Big(C_2 + \text{trust\_government} \cdot \text{trust\_science} \cdot \log(\text{age} + C_3)\Big) \cdot \text{risk\_of\_covid\_greater\_than\_vax} + K_1 \cdot \text{have\_covid\_sick\_family\_member} \cdot \log(\text{age} + C_4)$
Vaccinated (no booster)	$C_1 \cdot \text{covid\_threat} + C_2 \cdot \text{vaccine\_safe\_to\_me} + K_1 \cdot (\text{trust\_government} \cdot \text{trust\_science} \cdot \text{more\_attention\_to\_vax\_info} \cdot \sqrt{\text{age} + C_3})$
Booster	$C_1 \cdot e^{\text{age}^{C_2}} \cdot \text{covid\_threat} \cdot \sqrt{\text{vax\_protect\_long\_yes}} + C_3 \cdot \text{vaccine\_safe\_to\_me} + K_1 \cdot (\text{trust\_government} \cdot \text{trust\_science} \cdot \text{nurse} \cdot \sqrt{\text{age} + C_4})$

Feature: Age 18–38, income above county median.

Key take-aways for domain experts

- **Risk trade-off in vaccination choice.** The product covid_threat × risk_of_covid_greater_than_vax captures a critical decision-making trade-off. Messaging must narrow this perceived risk gap, e.g., by emphasizing robust evidence on vaccine safety [124].
- **Booster demand rises steeply with age.** The factor $e^{\text{age}^{C_2}}$ generates a nonlinear age effect: as age increases, perceived vaccine benefit grows rapidly. This reflects age-associated increases in risk perception and vulnerabilities [125].
- **Prior belief and healthcare occupation.** The presence of vax_protect_long_yes and nurse occupation in the booster equation means emphasizing extended protection and occupation will push this group further along the vaccination ladder [126].
- **Trust is pivotal for vaccine uptake.** The multiplicative trust_government × trust_science term appears in every vaccinated utility, signalling that confidence in both institutions amplifies willingness [127].

Mode	Discovered symbolic utility
Unvaccinated	$K_1 \cdot \sqrt{\text{covid\_threat} \cdot (\text{risk\_of\_covid\_greater\_than\_vax} + \sqrt{\text{age}} \cdot \text{gender} + C_1)} \cdot ((\text{trust\_government} \cdot \text{trust\_science})^2 + C_2) + K_2 \cdot \text{more\_attention\_to\_vax\_info} - K_3 \cdot (\text{income\_below\_median} \cdot \text{have\_university\_degree} \cdot (\text{trust\_government} \cdot \text{trust\_science})) + C_3$
Vaccinated (no booster)	$K_1 \cdot (\text{vaccine\_safe\_to\_me} + \text{trust\_government} \cdot \text{trust\_science} \cdot \sqrt{\sqrt{\text{age} + C_1} + \text{income\_unknown} + C_2}) + K_2 \cdot \frac{\text{more\_attention\_to\_vax\_info}}{\text{less\_attention\_to\_vax\_info} + C_3} + C_4$
Booster	$K_1 (\text{have\_covid\_sick\_family\_member} + \text{physician} (\text{trust\_government} \cdot \text{trust\_science}) \cdot (\sqrt{\text{age} + C_1} + C_2) + \text{nurse} \cdot (\text{trust\_science} \cdot \sqrt{\text{age} + C_3})) + C_4$

Feature: Adults with varied trust, income, and education profiles.

Key take-aways for domain experts

- **Information attention as lever.** Positive weights on more_attention_to_vax_info indicate that engagement with vaccine information consistently increases uptake → interactive campaigns remain essential [128].
- **Nonlinear trust amplification.** The squared term $(\text{trust\_government} \cdot \text{trust\_science})^2$ highlights a super-additive effect → boosting both trust dimensions together disproportionately reduces hesitancy [129].

- **Education buffers income hesitancy.** The negative income effect is mitigated by education–trust interactions → higher education plus trust can offset low-income hesitancy, pointing to education-focused outreach [130].

B Baseline Setup

B.1 Utility-Based Models

Table 9: Utility-based models and key settings (train : test = 0.8 : 0.2)

Model	Key (Non-default) Settings
SimpleMNL	`intercept="item"`; `optimizer="adam"`
ConditionalLogit	`optimizer="adam"`; added intercept for items 1 & 2
Latent Class MNL	`n_latent_classes=2`; `fit_method="mle"`; `optimizer="adam"`; `epochs=1000`

B.2 Machine Learning Models

Table 11: Machine learning models and key settings (train : test = 0.8 : 0.2)

Model	Best hyper-parameters
Logistic Regression	`C=10, penalty=l2, solver=saga`
Random Forest	`bootstrap=False, max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=600`
XGBoost	`colsample_bytree=0.8, learning_rate=0.05, max_depth=6, n_estimators=500, subsample=0.8`

B.3 LLM-Based Models

Take **Swissmetro** dataset as an example.

Prompt B.1: *Swissmetro* - Zeroshot

[SYS] You are a decision assistant that predicts a probability distribution over three travel modes, Swissmetro, Train, and Car, for a single trip.
You will receive two blocks of text:
<TRIP_INFO> ... details like trip purpose, luggage, payment, origin, destination ... </TRIP_INFO>
<TRANSPORT_OPTIONS> ... list of modes with travel time, cost, headway ... </TRANSPORT_OPTIONS>
Instructions:
1. Use only the information in <TRIP_INFO> and <TRANSPORT_OPTIONS>.
2. Estimate a probability for each mode so they sum to 1.
3. **Output only** a JSON object, for example:
“‘json { "Swissmetro": <float between 0 and 1>, "Train": <float between 0 and 1>, "Car": <float between 0 and 1> } “‘
No additional text; just the JSON object with normalized probabilities.
[USR] <TRIP_INFO> {trip_info} </TRIP_INFO>
<TRANSPORT_OPTIONS> {transport_options} </TRANSPORT_OPTIONS>

Prompt B.2: *Swissmetro* - Zeroshot-CoT

[SYS] You are a decision assistant that predicts a probability distribution over three travel modes, Swissmetro, Train, and Car, for a single trip.
You will receive two blocks of text:
<TRIP_INFO> ... details like trip purpose, luggage, payment, origin, destination ... </TRIP_INFO>
<TRANSPORT_OPTIONS> ... list of modes with travel time, cost, headway ... </TRANSPORT_OPTIONS>
Instructions:
1. Use only the information in <TRIP_INFO> and <TRANSPORT_OPTIONS>.
2. Estimate a probability for each mode so they sum to 1.
3. **Output only** a JSON object, for example:
“‘json { "Swissmetro": <float between 0 and 1>, "Train": <float between 0 and 1>, "Car": <float between 0 and 1> } “‘
No additional text; just the JSON object with normalized probabilities.
Let’s think step-by-step.
[USR] <TRIP_INFO> {trip_info} </TRIP_INFO>
<TRANSPORT_OPTIONS> {transport_options} </TRANSPORT_OPTIONS>

Prompt B.3: *Swissmetro* - Fewshot

[SYS] You are a decision assistant that predicts a probability distribution over three travel modes—Swissmetro, Train, and Car—for a set of travel records.
You will receive multiple records. Each record consists of three blocks: <TRIP_INFO> ... trip details: purpose, luggage, payment, origin, destination ... </TRIP_INFO> <TRANSPORT_OPTIONS> ... each mode's travel time, cost, headway ... </TRANSPORT_OPTIONS> <CHOICE> ... either a JSON object with probabilities (for examples), or left empty for the record to predict ... </CHOICE>
Instructions: - For records where <CHOICE> is filled, treat them as examples. - For the final record (with an empty <CHOICE>), output **only** the JSON object of normalized probabilities (summing to 1), with no extra text.
[USR] <TRIP_INFO> {trip_info_1} </TRIP_INFO> <TRANSPORT_OPTIONS> {transport_options_1} </TRANSPORT_OPTIONS> <CHOICE> {choice_1} </CHOICE> <TRIP_INFO> {trip_info_2} </TRIP_INFO> <TRANSPORT_OPTIONS> {transport_options_2} </TRANSPORT_OPTIONS> <CHOICE> {choice_2} </CHOICE> <TRIP_INFO> {trip_info_3} </TRIP_INFO> <TRANSPORT_OPTIONS> {transport_options_3} </TRANSPORT_OPTIONS> <CHOICE> {choice_3} </CHOICE> <TRIP_INFO> {trip_info_4} </TRIP_INFO> <TRANSPORT_OPTIONS> {transport_options_4} </TRANSPORT_OPTIONS> <CHOICE> {choice_4} </CHOICE> <TRIP_INFO> {trip_info_5} </TRIP_INFO> <TRANSPORT_OPTIONS> {transport_options_5} </TRANSPORT_OPTIONS> <CHOICE> {choice_5} </CHOICE> <TRIP_INFO> {trip_info_6} </TRIP_INFO> <TRANSPORT_OPTIONS> {transport_options_6} </TRANSPORT_OPTIONS> <CHOICE> Please predict the travel mode for this trip. </CHOICE>

Prompt B.4: *Swissmetro* - TextGrad

[INITIAL FULL PROMPT + SOLUTION] Task: Estimate the probability distribution over three travel modes (Swissmetro, Train, Car) for a single trip.
<TRIP_INFO> {trip_info} </TRIP_INFO>
<TRANSPORT_OPTIONS> {transport_options} </TRANSPORT_OPTIONS>
Solution (JSON): {"Swissmetro": 0.333, "Train": 0.333, "Car": 0.334}
[GRADING PROMPT] You are a transport-economics expert. Given the trip info, transport options, and predicted probabilities in the user's message, output a single line ONLY: Score: <float between 0 and 1> 1 = probabilities look highly reasonable, 0 = implausible. Remember: THE PREDICTION MUST BE A JSON DICT.

C Additional Experiments: Reasoning LLMs and End-to-End Baselines

Purpose and setup. This section probes how much backbone model capacity matters on our tasks. For a controlled comparison, we randomly sample 100 individuals from the 500-person pool to form a compact evaluation subset (same preprocessing, metrics, and decoding settings as in the main experiments). For each individual, we randomly sample one record. We evaluate ATHENA with five backbones: two state-of-the-art open-source reasoning models (`Qwen3-32B`, `DeepSeek-R1-Distill-Qwen-32B`) and three leading commercial offerings (`GPT-4o-mini`, `GPT-4o`, `Gemini-2.0-Flash`). Across both tasks, ATHENA attains *state-of-the-art classification performance* among LLM-based methods—consistently delivering the highest *Accuracy* and *F1*, with *AUC* that is competitive or superior to prompt-only LLM baselines (see Tables 13 and 1).

Structure dominates model size; stronger reasoning yields modest, consistent gains. Under ATHENA, swapping `GPT-4o-mini` for larger "reasoning" backbones (e.g., `GPT-4o`, `Qwen3-32B`, `DeepSeek-R1`) yields *incremental* but *consistent* improvements, especially on the more interaction-heavy *Vaccine* task. The effect is smaller on *Swissmetro*, where dominant explanatory factors (time/cost) are already well captured by the *symbolic discovery* $\rightarrow$ *textual refinement* pipeline. Intuitively, Stage 1 constrains the hypothesis space to interpretable utility forms, and Stage 2 makes small, directed edits to those forms; this turns the problem into guided search plus local adjustments. As a result, *structural bias* (symbolic utility discovery + semantic adaptation) shoulders most of the lift, while *backbone capacity* primarily fine-tunes edge cases (nonlinear interactions, atypical profiles), producing a steady but not dramatic gain.

Prompt-only methods are brittle and poorly calibrated; ATHENA regularizes both decisions and probabilities. Zero-shot / CoT / Few-shot prompting shows visible volatility across metrics: *Accuracy/F1* can spike on one dataset yet drop on another, and *AUC/CE* often swing with decoding details (temperature, sampling count, score-to-probability mapping). ATHENA markedly reduces this variance: the symbolic stage enforces cross-person consistency (shared operators, shared concept library), while the textual refinement stage adjusts *within* those constraints, leading to better class separability and more conservative probability mass. Empirically this manifests as stronger and more stable *F1/AUC*, with *CE* reflecting improved calibration compared to prompt-only baselines. In short, structure acts as *regularization* for both decisions and confidence.

End-to-end baselines trail on interpretability and robustness; ATHENA's decomposition captures heterogeneity with explicit utility logic. Machine learning-based models can be competitive on single metrics in isolated settings, but they do not expose explicit, policy-relevant utility functions and are less consistent across tasks/splits. They must implicitly learn both *which* attributes matter and *how* they combine, from scratch. ATHENA instead *decouples* the problem: Stage 1 discovers globally interpretable utility structure (operators, interactions), and Stage 2 adapts those structures to individual semantics. This yields *(i)* stronger across-task consistency in *Accuracy/F1/AUC*, *(ii)* end-to-end interpretability of the discovered utilities.

Table 13: Performance comparison across methods on *Swissmetro* and *Vaccine* datasets.

Method	**LLM Model**	**Swissmetro**				**Vaccine**			
		Acc.↑	F1↑	CE↓	AUC↑	Acc.↑	F1↑	CE↓	AUC↑
Zeroshot	gemini-2.0-flash	0.5800	0.3046	0.9059	0.6829	0.6000	0.5386	0.8317	0.7433
	GPT-4o-mini	0.6100	0.2763	0.9253	0.5556	0.5500	0.5302	0.8271	0.7500
	GPT-4o	0.5900	0.3310	0.8646	0.6946	0.6000	0.5465	0.8052	0.7306
	Qwen3	0.5900	0.4158	1.4047	0.6126	0.5400	0.5729	0.8676	0.7519
	DeepSeek-r1	0.5900	0.4339	1.2473	0.6608	0.6300	**0.6531**	0.8244	0.7692
Zeroshot-CoT	gemini-2.0-flash	0.5300	0.2809	0.9858	0.6409	0.6100	0.5485	0.8128	0.7604
	GPT-4o-mini	0.6100	0.2763	0.9109	0.6156	0.5900	0.5820	0.8161	0.7714
	GPT-4o	0.5800	0.3162	0.9237	0.6404	0.6300	0.5717	0.7785	0.7700
	Qwen3	0.6200	0.4632	1.7101	0.6284	0.5400	0.5843	0.9073	0.7364
	DeepSeek-r1	0.5800	0.3018	0.9522	0.6173	0.5200	0.5492	0.8761	0.7373
Few-shot	gemini-2.0-flash	0.7200	0.6922	10.0922	0.7984	0.5300	0.5508	14.2531	0.6747
	GPT-4o-mini	0.6800	0.5320	5.0402	0.7516	0.5200	0.5278	7.7066	0.6975
	GPT-4o	0.7300	0.6654	3.9386	0.8423	0.5700	0.5967	5.6261	0.7467
	Qwen3	0.7400	0.6760	7.0435	0.8033	0.5400	0.5487	9.5013	0.7060
	DeepSeek-r1	0.7000	0.6282	3.6154	0.8439	0.5500	0.5392	8.8895	0.6855
TextGrad	gemini-2.0-flash	0.5400	0.2432	1.1934	0.4718	0.5000	0.4511	4.1290	0.7345
	GPT-4o-mini	0.5700	0.3111	0.9551	0.5292	0.5000	0.4686	4.7960	0.6460
	GPT-4o	0.5600	0.3316	0.9441	0.6256	0.5600	0.5321	2.3468	0.6721
	Qwen3	0.5100	0.3669	2.0180	0.5322	0.4600	0.4276	6.4994	0.6303
	DeepSeek-r1	0.5800	0.3356	0.9631	0.6344	0.4300	0.4235	3.1080	0.5948
ATHENA (ours)	gemini-2.0-flash	**0.7900**	0.7185	**0.6121**	**0.9153**	0.6500	0.5978	0.8305	0.7998
	GPT-4o-mini	0.7600	**0.7304**	1.4208	0.8577	0.6500	0.6079	0.8034	0.8133
	GPT-4o	0.7700	0.7085	1.0417	0.8697	**0.6700**	0.6213	**0.7765**	**0.8279**
	Qwen3	0.7400	0.7040	4.9132	0.7754	0.5700	0.5650	1.1393	0.7637
	DeepSeek-r1	0.7100	0.6612	0.8437	0.8353	0.6600	0.6501	0.8115	0.8212

D Empirical Scalability Evidence

We benchmarked wall-clock time and token usage using `gpt-4o-mini` on the Swissmetro subset.

D.1 Stage 2 – Individual adaptation

Table 14: Runtime and token usage for Stage 2 (individual-level semantic adaptation) under different iteration counts T'.

T'	Time (s)	s/iter	tokens/iter
1	48.16	48.16	1079.6
3	176.89	58.96	1195.83
5	315.66	63.13	1249.28

D.2 Stage 1 – Group-level discovery

Table 15: Runtime and token usage for Stage 1 (group-level symbolic utility discovery) under different iteration counts T.

T	Time (min)	tokens total
5	30.61	215,281
15	36.69	251,974
30	65.64	479,751

These results confirm that runtime and token usage scale approximately linearly with the number of iterations, consistent with the theoretical analysis.

E Prompts

Take **Swissmetro** dataset as an example.

Prompt E.1: *Swissmetro* - Symbolic Utility Initialization

Step 1:
[SYS] You are a transportation planner specializing in analyzing the relationships among various factors that influence travel behavior. You will be provided with two types of information: individual features (delimited by <FEATURES> and </FEATURES>) and preliminary travel mode knowledge (delimited by <KNOWLEDGE> and </KNOWLEDGE>). Your task is to carefully review these inputs and in detailed sentence describe how the provided features interrelate. Ensure your response includes as many specific details as possible about the relationships, but do not propose any new features or suggest modifications to the existing ones. Example: Time: quadratic, Cost: log, luggage: linear.
[USR] <GROUP DESCRIPTION>{description}</GROUP DESCRIPTION> <FEATURES>{features}</FEATURES> <KNOWLEDGE>{knowledge}</KNOWLEDGE>
You should ONLY provide the relations between the features. YOU MUST return your assumption in this exact format: "'["relation_0","relation_1", ...]
Step 2:
[SYS] You are a helpful assistant that proposes mathematical expressions based on some provided suggestions. Your goal is to:
0. **Task**: Generate utility functions for travel mode choice of group of {description}.
1. **Use only** the specified variables: {variables}
2. **Represent all constants** with the symbol "C", and all coefficients with the symbol "K".
3. **Restrict** yourself to the following operators: operators
4. **For each group**, suggest utility functions for train, car, and Swissmetro respectively.
Your response must: - Propose exactly **{N}** groups of expressions. - MUST return in this exact format: "'[("expressions_car","expressions_train","expressions_metro"), ...]"', replace expressions_mode with your proposed expressions.
[USR] Suggestions: {suggestions}

Prompt E.2: *Swissmetro* - Results Analysis

[SYS] You are a creative and insightful mathematical research assistant. You have been provided with two sets of utility expressions: one function group labeled "Good Expressions" and one labeled "Bad Expressions." Your objective is to hypothesize about the underlying assumptions or principles that might generate the good expressions yet exclude the bad ones.
Key Points:
1. Focus primarily on the good expressions' mathematical structures and any connections they might have to physical or applied contexts.
2. Capital "C" in any expression is just an arbitrary constant.
3. Do not discuss or compare the expressions in terms of their simplicity or complexity.
4. Provide your reasoning step by step, but keep it very concise and genuinely insightful. No more than 5 lines.
[USR] Good Expression 1: (train: {texpr1}, car: {cexpr1}, metro: {mexpr1}), accuracy: {acc1}
Good Expression 2: (train: {texpr2}, car: {cexpr2}, metro: {mexpr2}), accuracy: {acc2}
Bad Expression 1: (train: {bexpr1}, car: {bexpr2}, metro: {bexpr3}), accuracy: {acc3}
Above expressions are travel mode choice utility functions of group of {description}. Propose {N} hypotheses that would be appropriate given the expressions. Provide short commentary for each of your decisions. Do not talk about topics related to the simplicity or complexity of the expressions. I want ideas that are unique and interesting enough to amaze the world's best mathematicians.

Prompt E.3: *Swissmetro* - Crossover

[SYS] You are a helpful assistant that recombines two mathematical expressions based on some provided suggestions. Your goal is to produce new expressions that:
1. Blend or merge elements from both reference expressions in a way that reflects the suggestions.
2. Adhere to the following constraints:
- You may only use the variables in library: {variables}
- All constants must be represented with the symbol C
- Only the following operators are allowed: {operators}
Guidelines:
- Propose exactly {N} new expressions.
- Each new expression should integrate elements of both reference expressions. You can also propose new terms with variables that are in the library but not in the old expressions.
- If any suggestions appear contradictory, reconcile them reasonably.
MUST return in this exact format:
`[("expressions_car","expressions_train","expressions_metro"), ...]`
Replace `expressions_` with your proposed expressions.
[USR] Suggestion: {suggestions}
Reference Expression group 1: (train: {texpr1}, car: {cexpr1}, metro: {mexpr1})
Reference Expression group 2: (train: {texpr2}, car: {cexpr2}, metro: {mexpr2})

Propose {N} expressions that would be appropriate given the suggestions and references.

Prompt E.4: *Swissmetro* - Mutation

[SYS] You are a helpful assistant that generates mutated variants of a **triplet** of mathematical expressions (car, train, metro) based on provided mutation strategies. Your goal is to produce new expression triplets that: 1. Mutate the reference expressions by applying mutation operations (e.g., adjust coefficients, swap variables, alter operators) in a way that reflects the suggestions. 2. Adhere to the following constraints: - You may only use the variables in library: {variables} - All constants must be represented with the symbol C - Only the following operators are allowed: {operators}
Guidelines: - Produce exactly {M} mutated **triplets**. - Within each triplet you must provide one mutated expression for **car**, one for **train** and one for **metro**. - A mutation can modify any combination of variable, operator or constant, but each expression must remain syntactically valid under the constraints.
MUST return in this exact format:
`[("mut_car1","mut_train1","mut_metro1"), ...]`
[USR] Generate {M} mutated variants of the following mathematical expression triplet according to these mutation strategies: – You may only use variables from: {variables} – All constants must be written as C – Only these operators are allowed: {operators}
Mutation strategies: {suggestions}
Reference expressions: (car): {cexpr} (train): {texpr} (metro): {mexpr}
Please return exactly {M} new, syntactically valid triplets in the JSON list format shown above.

Prompt E.5: *Swissmetro* - Semantic Adaptation Initialization

[SYS] You are a travel-behavior preference selector. You will be given two blocks of information:
<DEMOGRAPHICS> ... </DEMOGRAPHICS> <UTILITY_FUNCTION> ... </UTILITY_FUNCTION>
Your goal: choose the single best-matching high-level preference template for this group **exactly** from the catalogue below and output **only** the template name (uppercase).
CATALOGUE - TIME_EFFICIENCY : travellers primarily minimise total travel time.
- COST_SAVING : travellers primarily minimise direct monetary cost.
- COMFORT_SEEKING : travellers value comfort/service frequency and dislike crowding.
- BALANCED : sensitivities are evenly distributed across factors.
...//OTHER POSSIBLE TEMPLATE
Return nothing else — no commentary, no punctuation, just the template name.
[USR] <DEMOGRAPHICS>{demographics}</DEMOGRAPHICS>
<UTILITY_FUNCTION>{utility}</UTILITY_FUNCTION>

Prompt E.6: *Swissmetro* - Semantic Adaptation Loss Function

Evaluate the travel mode prediction based on the individual's profile and alternatives. Compare it to the actual choice and identify any discrepancies. Be concise and focus on why the prediction might be incorrect. Return 0 if they match, 1 otherwise.

Prompt E.7: *Swissmetro* - Prediction

[SYS] You are a decision assistant that recommends the most suitable travel mode for an individual trip by estimating a probability distribution over three options: Swissmetro, Train, and Car.
You will receive three blocks: <TEMPLATE> ... optimized preference template ... </TEMPLATE>
<PROFILE> ... individual profile ... </PROFILE>
<ALTERNATIVES> ... attributes of Swissmetro, Train, and Car ... </ALTERNATIVES>
Instructions: 1. **Use the <TEMPLATE> as a guide** for understanding the individual's likely preference bias (e.g., time efficiency, cost saving, comfort seeking, balanced).
2. **Analyze the <PROFILE>** (age, gender, income, trip details) **and the <ALTERNATIVES>** (travel time, cost, headway).
3. **Estimate and output a probability** for each travel mode, such that all three probabilities sum to 1.
Output format (JSON only): "'json { "Swissmetro": <float between 0 and 1>, "Train": <float between 0 and 1>, "Car": <float between 0 and 1> } "' No additional text; just the JSON object with normalized probabilities.
[USR] <TEMPLATE>{template_name}</TEMPLATE> <PROFILE> {individual_block} </PROFILE> <ALTERNATIVES> {options} </ALTERNATIVES>