LEARNING A FOUNDATION MODEL OF HUMAN SOCIAL REASONING FROM HETEROGENEOUS EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have demonstrated strong capabilities in symbolic reasoning, tool use, and complex task solving, yet they still lack a fundamental understanding of human social reasoning: how people interpret the same situation, weigh relevant information, and arrive at decisions. This limitation largely stems from the scarcity of explicit decision knowledge in pretraining corpora about how individual information leads to behavioral choices. To recover such knowledge from structured behavioral data, we use heterogeneous experts with complementary inductive biases to uncover relationships between individual information and behavioral choices. Building on these relationships, we introduce **SocialReasoner** to supplement LLMs with the decision knowledge required for human social reasoning through natural language reasoning supervision. Drawing on four major survey programs, we construct nine large datasets for human social reasoning, comprising 281,995 respondents and over 8.7 million observed responses across 365 questions, 13 social domains, and diverse countries and regions. Across multiple model families and scales, **SocialReasoner** consistently outperforms all evaluated state-of-the-art baselines, with average relative accuracy gains exceeding 40% and substantially better alignment with observed population behavioral distributions. It further generalizes to an unseen survey and extends to decision tasks in finance and healthcare. Code and models will be made publicly available.

1 INTRODUCTION

Recent advances in large language models (LLMs) have demonstrated strong capabilities in reasoning (Jaech et al., 2024; Guo et al., 2025; Team, 2026), scientific question answering (Rein et al., 2024), mathematics (Yang et al., 2024), and code generation (Zheng et al., 2025; Zhang et al., 2026). However, these capabilities do not imply an understanding of how humans reason about social situations and make decisions.

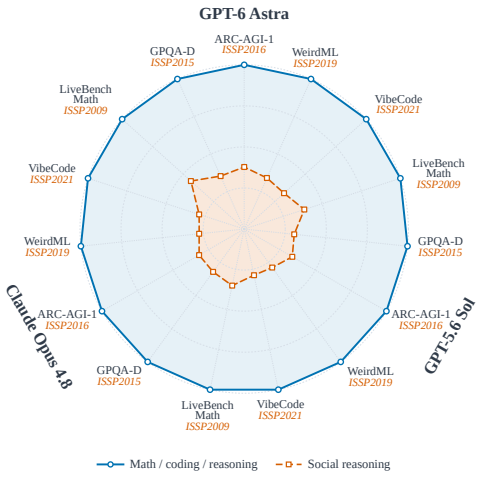

Figure 1: **State-of-the-art LLMs show strong general reasoning but limited human social reasoning.**

For example, prior studies show that LLMs substitute user agreement for truthfulness (Sharma et al., 2024), homogenize diverse human responses (Wang et al., 2025a), and suppress human uncertainty by producing unrealistically accurate factual estimates (Aher et al., 2023). Our evaluation further illustrates this gap: leading commercial LLMs excel on mathematics, science, and coding benchmarks, yet remain limited in modeling individual human decisions (Fig. 1). Although scaling model capacity has driven broad progress in LLMs (Kaplan et al., 2020), the persistence of this gap even among leading commercial models suggests that scaling capacity alone may be insufficient. Recent work similarly finds uneven scaling in social reasoning, particularly for tasks involving human biases (Ziems et al., 2026). Our analysis similarly

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

shows that the relationship between model scale and social reasoning performance weakens among larger models (Appendix A.7). Together, these findings motivate looking beyond scale toward the knowledge required for human social reasoning. Unlike mathematics and coding tasks, much of this knowledge is not explicitly expressed in natural language corpora, but remains implicit in structured survey and experimental records that document individual information and behavioral choices. Although widely available, these records rarely reveal how such information leads to decisions. A central scientific question thus emerges: *How can LLMs model human decision making and the underlying social reasoning processes from structured behavioral data?*

Answering this question requires addressing two linked challenges. First, we need to identify which individual factors shape each behavioral choice and how they relate to the resulting decision. Second, we need to transform these decision relationships into a form that LLMs can learn from to model the underlying human reasoning process and decision logic.

The first challenge lies in how human behavioral data represent individuals. Existing datasets commonly characterize a person primarily through demographic attributes and use these profiles to model individual choices (Santurkar et al., 2023; Chen et al., 2024; Wang et al., 2025b). However, demographic information mainly describes who a person is, rather than how that person perceives and evaluates a particular issue. Recent work further emphasizes that modeling human thought is essential for understanding human behavior (Li et al., 2026a), suggesting that representations of how individuals think may be more informative than demographic profiles alone. Empirically, we find that incorporating perception information substantially improves individual decision prediction even without additional training, whereas removing demographic attributes causes only modest performance degradation (Fig. 2(a)). Structural equation modeling (Bollen, 2014) further shows that perception factors are more strongly associated with behavioral choices than demographic factors (Fig. 2(b)). Motivated by these findings, we return to the original survey records and reconstruct nine large-scale datasets for human social reasoning from the American Trends Panel (Pew Research Center, 2024a;b), European Values Study (EVS, 2022), International Social Survey Programme (ISSP Research Group, 2017a;b; 2018; 2022; 2024), and General Social Survey (Davern et al., 2024). In these datasets, individuals are represented jointly by demographic attributes and perception measurements, providing richer behavioral representations for modeling human decision making. Dataset construction details are provided in Appendix E.

With richer individual representations in place, a second challenge remains: identifying which information matters for a particular decision and how it shapes the resulting choice. As shown in Fig. 2(c), providing more perception information does not continuously improve performance, suggesting that richer information alone is insufficient. This limitation cannot be overcome simply by stronger general reasoning: large reasoning models (LRMs) remain limited in modeling human decisions (Appendix A.6).

To address these challenges, we introduce **SocialReasoner**, which learns human social reasoning from structured behavioral data by integrating heterogeneous experts with distinct inductive biases. These experts uncover complementary relationships between individual information and behavioral choices. The framework then transforms these relationships into natural language reasoning supervision, allowing LLMs to learn how relevant individual signals support different behavioral choices. In this way, SocialReasoner provides a shared foundation for modeling human social reasoning across heterogeneous decisions. Across datasets spanning diverse domains of human decision making and social cognition, from technology attitudes and institutional trust to social inequality, work values, and social cohesion, the proposed method consistently outperforms frontier LLMs such as GPT-6 Astra (OpenAI, 2026b) and behavioral adaptation baselines such as SAFO (Lin et al., 2026) and HARPO (Ong et al., 2026). To summarize, our contributions are as follows:

- We identify two fundamental bottlenecks in modeling human social reasoning from structured behavioral data. First, current approaches do not explicitly identify which individual factors shape a particular behavioral choice. Second, they lack a mechanism to transform these decision relationships into knowledge that LLMs can learn from to model the underlying human reasoning process and decision logic.

- To address these challenges, we reconstruct nine datasets for human social reasoning from four major survey programs and propose **SocialReasoner**, which integrates heterogeneous experts with distinct inductive biases to uncover complementary decision relationships

from structured behavioral data and transfer them to LLMs through natural language reasoning supervision.

- Extensive experiments across multiple model families and scales show that **SocialReasoner** consistently improves human social reasoning. At the population level, its aggregated predictions more faithfully recover observed patterns of human behavior. The method further generalizes to an unseen survey and transfers to decision tasks in finance and healthcare, demonstrating its effectiveness and generalizability.

2 Background and Motivation

In this section, to investigate what limits the human social reasoning capabilities of current LLMs and whether perception information provides a critical source of individual decision signals, we first introduce how our datasets are constructed (Section 2.1), establishing the empirical basis for studying richer individual representations and their relationships with behavioral choices. We then present the key observations derived from these datasets (Section 2.2), which motivate our subsequent modeling approach. Related work is discussed in Appendix D.

2.1 Dataset Construction

Our datasets are reconstructed from large scale surveys in which each respondent answers a broad set of questions. Beyond demographic attributes, these surveys contain responses that reflect individual attitudes, beliefs, values, trust, preferences, experiences, and self assessments, providing richer information about how individuals perceive and evaluate social issues and their own circumstances. We collectively refer to these responses as *perception information*.

For respondent u_i, let d_i denote the demographic attributes, and let

$$p_i = \{(q_k, y_{i,k})\}_{k \in \mathcal{P}_i} \tag{1}$$

denote the perception profile, where $\mathcal{P}_i$ indexes the non-demographic survey items answered by u_i that are available for constructing the profile, q_k denotes the corresponding survey item, and $y_{i,k}$ denotes the observed response of u_i to that item. Together, these directly observed responses provide a richer individual representation by capturing how respondents perceive and evaluate social issues beyond what demographic attributes alone can describe.

Based on the perception profile p_i, we construct prediction instances by selecting one survey item answered by u_i as the target question and using the remaining information to predict the respondent's choice. Specifically, for a target question q_j with candidate options $\mathcal{O}_j$, we construct the corresponding input as

$$x_{i,j} = \left(d_i,\, p_i^{(-j)},\, q_j,\, \mathcal{O}_j\right), \tag{2}$$

where $p_i^{(-j)}$ denotes the perception profile with the target question q_j and its observed response excluded. The observed response $y_{i,j}$ represents the resulting behavioral choice. With this construction, our dataset setting instantiates the human decision-making process: d_i and $p_i^{(-j)}$ characterize the individual, q_j and $\mathcal{O}_j$ define the current decision context, and $y_{i,j}$ represents the resulting behavioral choice. The underlying social reasoning process through which individuals interpret and weigh this information before arriving at a decision remains unobserved, motivating our study of how it can be learned from structured behavioral data.

2.2 Key Observations

To understand what information is needed to model this underlying social reasoning process, we first examine whether perception information provides decision-relevant signals beyond demographic attributes. Using the datasets constructed above, we further analyze how these signals relate to behavioral choices and whether current LLMs can effectively use them.

Perception Provides Stronger Decision Signals. As shown in Fig. 2(a), incorporating perception information substantially improves the modeling of individual choices across models and datasets

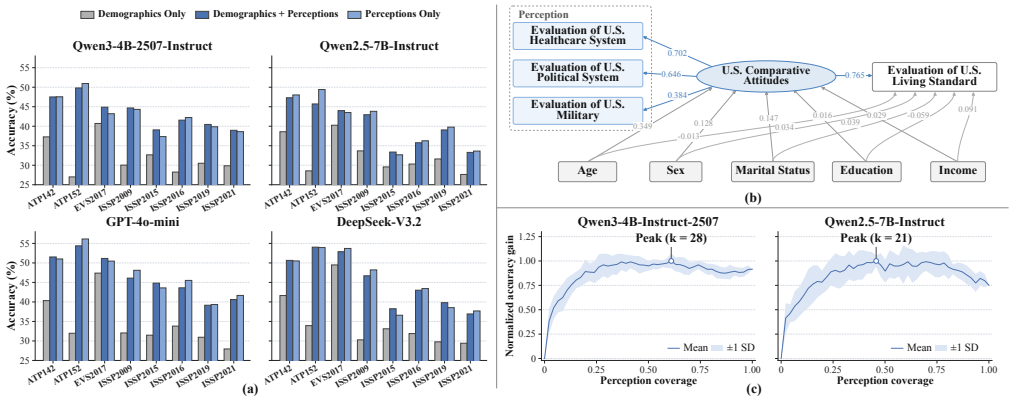

Figure 2: **Perception provides stronger decision signals, but more is not always better.** (a) Removing perception information causes substantially greater performance degradation than removing demographic attributes. (b) Structural equation modeling shows that perception factors are more strongly associated with behavioral choices than demographic factors. (c) Adding more perception information yields non-monotonic gains, with performance peaking before all available perceptions are included.

without additional training, whereas removing demographic attributes causes only modest performance degradation. This suggests that perceptions contain decision-relevant information beyond demographic background. To further examine how such information relates to human behavior, we conduct structural equation modeling (SEM). Fig. 2(b) shows that perception factors exhibit substantially stronger associations with behavioral choices than demographic attributes. Together, these results indicate that perception information provides a richer basis for modeling human decision making. Detailed analyses are provided in Appendix A.2.

More Perception Is Not Always Better. We further examine how model behavior changes as progressively more perception information becomes available. Starting from no perception information, we increase the number of included perception items until the full profile is provided. For each intermediate number of items, we randomly sample 10 perception subsets and report the resulting mean performance and variation; the number of included items is normalized by the total available perceptions for visualization. As shown in Fig. 2(c), performance improves rapidly at first, but peaks before all available perceptions are included and then declines. This non-monotonic pattern suggests that pretrained LLMs can exploit broad regularities in human perceptions, but cannot reliably distinguish which signals matter for a particular decision or how they should be combined. The bottleneck therefore lies not in access to richer individual information, but in learning the decision relationships that connect relevant perceptions to behavioral choices.

Together, these observations show that perception information is necessary but not sufficient for modeling human decision making: models need richer individual signals, but must also learn which signals matter for each decision and how they shape behavioral choices.

3 Method

3.1 Overview

Motivated by the observations above, we introduce **SocialReasoner** to model the social reasoning underlying human decisions from structured behavioral data. As shown in Fig. 3, the framework first leverages heterogeneous experts with distinct inductive biases to identify complementary factors and relationships that shape behavioral choices (Section 3.2). These relationships are then made explicit and transformed into natural language reasoning supervision that captures how individual information leads to human decisions (Section 3.3). Finally, the diverse reasoning patterns recovered from heterogeneous experts are jointly distilled into a unified LLM for human social reasoning (Section 3.4).

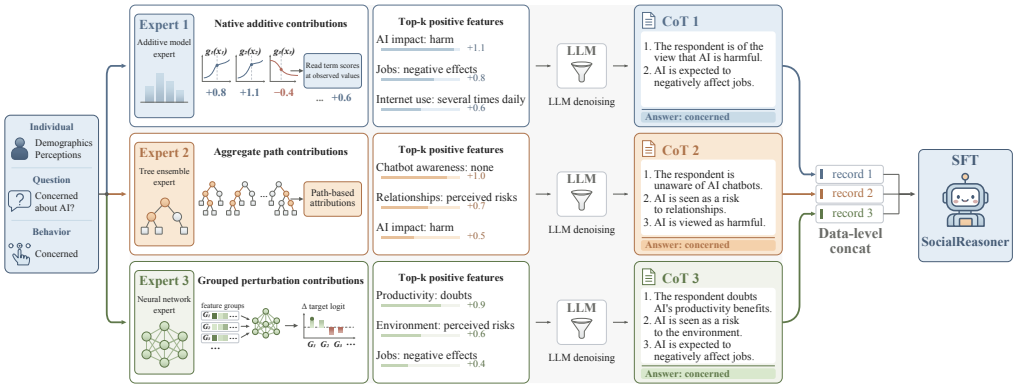

Figure 3: **Overview of SocialReasoner.** SocialReasoner leverages heterogeneous experts to uncover complementary relationships between individual information and behavioral choices, and distills these relationships into LLMs for human social reasoning.

3.2 Heterogeneous Decision Experts

We use heterogeneous experts to uncover relationships between individual information and behavioral choices. For each target question q_j, an expert models the observed choice $y_{i,j}$ from d_i and $p_i^{(-j)}$. Because each question represents a distinct decision problem with its own choice space, we learn a separate expert for each question.

To capture complementary decision relationships, we consider three modeling paradigms with distinct inductive biases: additive, tree-based, and neural models, instantiated by EBM, LightGBM, and MLP, respectively. We denote the set of experts as $\mathcal{M}$. Their different inductive biases allow them to capture complementary relationships underlying the same behavioral choice.

To avoid recovering decision relationships from instances used to fit the experts, we obtain predictions on held out folds. For respondent u_i in fold k, the prediction from expert m for question q_j is

$$\hat{y}_{i,j}^{(m)} = f_{j,-k}^{(m)}\left(d_i, p_i^{(-j)}\right), \tag{3}$$

where $f_{j,-k}^{(m)}$ is fitted on the remaining folds. We retain predictions satisfying $\hat{y}_{i,j}^{(m)} = y_{i,j}$, so that subsequent reasoning construction is grounded in relationships consistent with observed human behavior. At this stage, the experts provide only behavioral predictions; the underlying social reasoning process remains implicit.

3.3 Decision Evidence Extraction and Refinement

Local Decision Evidence. We make the relationships captured by each expert explicit by identifying the input factors that support its prediction. For each retained prediction, we derive local feature attributions $\mathbf{c}_{i,j}^{(m)}$ and select the factors with the largest positive contributions:

$$\mathbf{e}_{i,j}^{(m)} = \mathrm{TopK}^{+}\left(\left(d_i, p_i^{(-j)}\right), \mathbf{c}_{i,j}^{(m)}, K_e\right), \tag{4}$$

where TopK^{+} selects at most K_e input factors with the largest positive contributions. The resulting $\mathbf{e}_{i,j}^{(m)}$ represents the candidate decision evidence provided by expert m. Because attribution is performed independently for each expert, different inductive biases can reveal complementary factors for the same decision. Detailed attribution procedures are provided in Appendix C.1.

LLM Semantic Denoising. Statistical importance does not necessarily imply semantic relevance to a human decision. A factor may contribute to an expert prediction through correlations in the data without providing a meaningful basis for the corresponding choice. We therefore use GLM-5.3-Flash (Zeng et al., 2026) to refine the candidate evidence with respect to the target question

q_j and candidate options $\mathcal{O}_j$. The model receives neither the observed choice $y_{i,j}$ nor the expert prediction $\hat{y}_{i,j}^{(m)}$, preventing the target decision from guiding the refinement. The resulting evidence $\widetilde{\mathbf{e}}_{i,j}^{(m)}$ retains the information used to construct the subsequent social reasoning. Details of semantic denoising and reasoning construction are provided in Appendix C.2.

3.4 Heterogeneous Decision Knowledge Distillation

The refined evidence from heterogeneous experts provides complementary views of how individual information relates to behavioral choices. We transform each expert's evidence into natural language reasoning that makes this relationship explicit. Rather than collapsing different expert views into a single explanation, we preserve them as separate supervision, allowing the LLM to learn from diverse reasoning patterns grounded in observed human behavior.

For expert m, let $o_{i,j}^{(m)}$ denote the supervision output consisting of the resulting social reasoning followed by the observed choice $y_{i,j}$. Following the output format of DeepSeek-R1 (Guo et al., 2025), the reasoning is enclosed by `<think>` and `</think>`. Let $\mathcal{D}^{(m)}$ denote the corresponding supervision set. We jointly distill the reasoning derived from all experts into the LLM through supervised fine-tuning:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_{m\in\mathcal{M}} \sum_{(x,o)\in\mathcal{D}^{(m)}} \sum_{t=1}^{|o|} \log P_\theta\left(o_t \mid x, o_{<t}\right), \tag{5}$$

where θ denotes the LLM parameters.

The refined evidence is used only to construct the reasoning supervision, whereas the LLM is trained on the complete input $x_{i,j}$. Thus, the experts expose relationships underlying behavioral choices, while the LLM learns to identify relevant information and generate both the social reasoning process and the resulting decision.

4 Experiment

4.1 Main Results

Comparison with Frontier LLMs. Table 1 compares SocialReasoner with frontier LLMs across eight datasets. Despite using substantially smaller 4B–8B backbones, SocialReasoner outperforms GPT-6 Astra, the strongest evaluated frontier model, by more than 4 percentage points on average and achieves higher performance on every dataset. These results demonstrate the value of learning human social reasoning from structured behavioral data beyond relying on general model capability alone.

Comparison with Non-LLM Baselines. We compare SocialReasoner with Random Guess, classical cognitive models (WADD, EW, TTB, and GCM), GEMS, and JEV. Among these methods, TTB is particularly competitive, outperforming GEMS on seven of the eight datasets and JEV on all eight. Nevertheless, SocialReasoner consistently surpasses TTB across all datasets and backbone families, with an average advantage of more than 4 percentage points. The gap is even larger over JEV, a recent probabilistic decision model, reaching approximately 11 percentage points on average. These results suggest that while compact cognitive rules can capture useful structure in human decisions, explicitly learning heterogeneous decision relationships provides substantially stronger support for human social reasoning. Details of the cognitive baselines are provided in Appendix A.1. We further analyze the performance and complementarity of different decision experts in Appendix B.2.

Effectiveness across Backbone Models. We compare SocialReasoner with representative methods from the major technical paradigms for modeling human behavior. For prompting methods, we include CoT, ClaimSim, and Few-Shot Neighbor. For supervised adaptation, we compare against standard SFT and SAFO, which is specifically designed for human behavioral data. For population response distribution alignment, we include SubPop. We further consider DPO as a baseline for preference optimization and HARPO as a reinforcement learning method designed for heterogeneous human behavior. Across all eight datasets, SocialReasoner consistently improves the corresponding base models, achieving relative gains in average accuracy of 34.37%, 45.22%, and 45.38%

Table 1: **Main results on human social reasoning.** Accuracy (%) across eight datasets and three backbone families.

Method	ATP142	ATP152	EVS2017	ISSP2009	ISSP2015	ISSP2016	ISSP2019	ISSP2021
Frontier LLMs								
Claude Opus 4.8 (Anthropic, 2026)	51.53	60.52	56.61	49.09	45.98	45.43	40.96	42.35
GPT-6 Astra (OpenAI, 2026b)	55.10	61.71	63.84	57.59	49.20	51.73	47.88	46.42
GPT-5.6 Sol (OpenAI, 2026a)	54.10	59.72	57.59	52.44	44.27	47.64	42.50	42.24
GPT-4o-mini (OpenAI, 2024)	51.53	54.38	51.15	46.11	44.82	43.62	39.16	40.62
DeepSeek-V4-Pro (Xu et al., 2026)	52.55	59.84	56.23	53.54	43.48	45.94	42.05	42.27
DeepSeek-V4-Flash (Xu et al., 2026)	54.59	56.32	57.13	52.66	43.16	47.23	44.06	42.24
DeepSeek-V3.2 (Liu et al., 2025a)	50.65	54.04	52.88	46.70	38.26	43.01	39.81	36.92
Gemma-4-31B (Team et al., 2026)	52.52	57.78	50.98	49.30	42.41	46.25	43.08	43.89
Llama-3.1-70B (Grattafiori et al., 2024)	51.21	52.33	55.03	46.54	42.94	44.63	41.83	42.39
Qwen3.5-Flash (Qwen Team, 2026)	51.93	56.08	52.10	52.38	41.68	45.06	42.34	37.98
Qwen3.6-27B (Team, 2026)	50.91	55.30	49.88	47.69	38.50	43.25	42.28	38.63
Qwen3.5-397B-A17B (Qwen Team, 2026)	52.55	59.21	54.82	52.49	44.18	46.82	42.53	41.19
GLM-5.3-Flash (Zeng et al., 2026)	51.53	59.64	55.83	53.58	44.67	48.31	43.46	41.91
Non-LLM Baselines								
Random Guess	24.25	22.56	24.69	16.98	17.68	19.06	19.86	20.85
JEV (Almeida, 2026)	50.51	52.19	56.04	52.84	41.05	46.15	39.26	40.48
GEMS (Suh et al., 2026)	53.10	58.27	57.03	50.65	50.90	49.49	49.03	50.05
WADD (Payne et al., 1988)	46.36	43.06	46.91	40.72	40.98	41.50	38.65	39.69
EW (Dawes, 1979)	46.35	42.77	46.84	40.69	40.53	41.41	38.58	39.46
TTB (Gigerenzer & Goldstein, 1996)	55.45	58.66	59.05	55.53	51.49	52.64	49.66	49.98
GCM (Nosofsky, 1986)	46.89	45.66	49.19	41.97	42.54	43.47	39.47	42.47
Qwen3-4B-2507-Instruct								
Baseline (Yang et al., 2025)	47.51	49.81	44.88	44.68	39.06	41.56	40.46	38.95
CoT (Wei et al., 2022)	47.45	52.08	47.24	42.38	33.76	40.65	37.60	34.53
Few-shot Neighbor (Hwang et al., 2023)	49.49	53.90	50.42	48.46	44.85	45.44	43.38	41.06
ClaimSim (Yu et al., 2025)	47.00	50.80	47.30	45.56	38.88	41.23	39.95	36.94
SubPop (Suh et al., 2025)	43.16	43.45	43.36	36.31	33.68	35.07	30.01	32.34
DPO (Rafailov et al., 2023)	54.33	55.19	61.23	56.53	50.01	52.94	49.61	45.46
SFT (Kolluri et al., 2025)	53.18	58.89	60.10	54.64	50.40	51.99	49.12	49.25
SAFO (Lin et al., 2026)	50.49	55.00	56.88	51.10	46.63	48.06	45.23	45.00
HARPO (Ong et al., 2026)	55.03	49.88	58.65	56.49	49.83	52.18	50.04	48.72
Ours	**58.53**	**63.10**	**64.16**	**59.34**	**55.71**	**56.89**	**53.91**	**54.49**
Qwen2.5-7B-Instruct								
Baseline (Team, 2024)	47.30	45.71	44.00	42.95	33.38	35.77	39.05	33.29
CoT (Wei et al., 2022)	45.33	47.58	43.91	44.03	32.97	36.43	38.70	35.05
Few-shot Neighbor (Hwang et al., 2023)	48.26	50.17	49.97	44.76	38.74	42.68	39.77	34.81
ClaimSim (Yu et al., 2025)	44.21	43.51	41.83	43.36	35.91	36.48	36.46	36.12
Subpop (Suh et al., 2025)	44.70	44.50	45.43	39.46	37.70	39.22	36.44	37.28
DPO (Rafailov et al., 2023)	54.69	52.14	50.81	56.67	43.88	43.91	44.34	38.21
SFT (Kolluri et al., 2025)	53.95	58.23	51.58	50.15	45.76	47.68	46.30	43.61
SAFO (Lin et al., 2026)	51.96	54.89	56.93	51.18	46.35	48.21	45.09	44.78
Ours	**58.27**	**63.22**	**64.12**	**58.19**	**58.18**	**55.80**	**54.39**	**54.63**
Llama-3.1-8B-Instruct								
Baseline (Grattafiori et al., 2024)	45.22	45.71	48.04	41.34	33.83	36.60	36.73	33.69
CoT (Wei et al., 2022)	45.26	42.81	48.55	37.38	28.93	35.19	36.79	29.95
Few-shot Neighbor (Hwang et al., 2023)	47.10	52.68	52.86	46.81	43.39	45.80	42.90	41.90
ClaimSim (Yu et al., 2025)	46.18	42.75	48.66	41.18	34.62	38.08	38.57	32.68
Subpop (Suh et al., 2025)	42.81	40.58	43.82	35.14	33.93	35.92	32.74	33.38
DPO (Rafailov et al., 2023)	52.32	53.78	57.34	56.18	49.46	52.43	50.09	49.09
SFT (Kolluri et al., 2025)	53.59	58.42	60.18	54.62	50.25	51.96	49.22	49.22
SAFO (Lin et al., 2026)	50.95	56.16	57.16	50.81	45.82	48.76	45.15	44.86
Ours	**58.16**	**62.88**	**64.28**	**59.41**	**56.51**	**56.68**	**54.49**	**54.50**

on Qwen3-4B, Qwen2.5-7B, and Llama-3.1-8B, respectively. These consistent gains across model families and technical paradigms demonstrate the effectiveness of learning human social reasoning from structured behavioral data.

Beyond individual behavioral fidelity, we examine whether these improvements also recover population patterns of human behavior. Figure 4(a) shows that SocialReasoner produces an aggregate response distribution that more closely matches the observed human distribution on ISSP2021. Importantly, SocialReasoner does not directly optimize or calibrate population distributions. Nevertheless, Figure 4(b) and the full results in Appendix B.1 show that it outperforms SubPop, which explicitly targets population response distribution alignment, across most datasets and evaluation metrics. These results suggest that more faithful population behavior can emerge from better modeling heterogeneous individual reasoning, rather than direct correction of aggregate response frequencies. One possible explanation is that heterogeneous experts provide diverse social reasoning supervision that better preserves variation across individuals, which in turn leads to more faithful population distributions.

Cross-Dataset Generalization. Using Qwen2.5-7B-Instruct, we train all methods on a randomly sampled 10% of the instances from the eight source datasets and evaluate on the entirely unseen GSS2022. As shown in Table 2(a), SubPop, SAFO, and HARPO all reduce overall accuracy rela-

Table 2: **Generalization across surveys and decision domains.** Accuracy (%) on (a) GSS2022 and (b) finance and healthcare tasks.

(a) Generalization across surveys: GSS2022

Backbone	Base	SubPop	SAFO	HARPO	Ours
Qwen2.5-7B-Instruct	49.85	40.53	49.70	43.82	**51.68**

(b) Applicability across domains

	Finance			Healthcare	
Method	**Credit-D**	**Credit-G**	**HomeLoan**	**AD**	**Diabetes**
Qwen3-4B-2507-Instruct					
Base	74.84	45.50	73.98	55.74	64.94
CoT	70.76	55.50	74.80	55.07	63.64
SFT	76.12	57.00	73.98	64.70	58.44
SAFO	75.52	49.00	73.98	57.94	68.83
DPO	78.06	58.00	73.98	60.64	71.43
GRPO	77.16	46.00	73.98	56.59	68.18
Ours	**78.66**	**74.00**	**76.42**	**67.06**	**78.57**
Qwen2.5-7B-Instruct					
Base	74.42	44.50	73.17	44.76	71.43
CoT	74.84	48.00	73.17	56.42	62.99
SFT	75.82	65.00	74.80	68.41	73.38
SAFO	75.82	69.00	74.80	68.41	73.38
DPO	76.29	67.00	73.98	48.65	74.68
GRPO	76.29	64.00	73.98	44.93	70.78
Ours	**78.56**	**73.00**	**77.24**	**70.78**	**77.27**

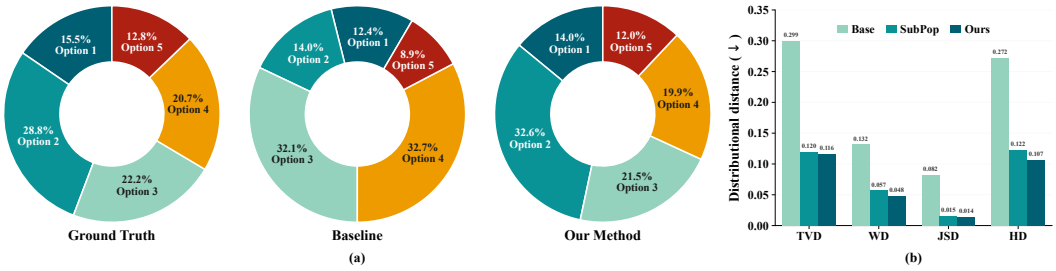

Figure 4: **Population-level behavioral patterns on ISSP2021. (a)** Aggregate response distributions with options harmonized into five categories. **(b)** Mean distributional distances from observed human responses across questions using their original response options. All model results use Qwen3-4B as the backbone.

tive to the base model, whereas SocialReasoner improves it from 49.85% to 51.68%, with gains in seven of the nine domains. In Status Attainment & Mobility, accuracy increases from 35.45% to 44.22%, corresponding to a relative gain of 24.74%. This contrast suggests that adapting to source behavioral patterns alone offers limited transfer, whereas learning the relationships between individual information and human decisions provides more generalizable knowledge. By transferring these relationships as social reasoning supervision, SocialReasoner generalizes to a new survey with different individual information, questions, and choice spaces.

Applicability Across Decision Domains. We further apply SocialReasoner to five decision tasks in finance and healthcare (Grinsztajn et al., 2022; Hegselmann et al., 2023; Konapure, 2023; Kwon et al., 2024). As shown in Table 2(b), the method achieves the best performance on all five tasks under both backbones, with relative gains of 18.96% and 22.24% over the corresponding base models. These results suggest that the proposed framework extends beyond human social reasoning to settings where decision processes remain implicit in structured data that record individual information and behavioral outcomes.

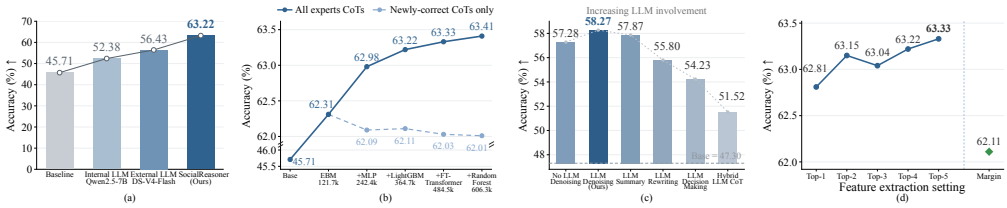

Figure 5: **(a)** Sources of social reasoning supervision on ATP152. **(b)** Reasoning diversity from heterogeneous experts on ATP152. **(c)** Decision grounding and LLM involvement on ATP142. **(d)** Selection of decision-relevant signals on ATP152. All experiments use Qwen2.5-7B-Instruct as the backbone.

4.2 Ablation Studies and In-Depth Analysis

Heterogeneous Experts Capture Diverse Social Reasoning. We first compare reasoning traces generated by Qwen2.5-7B-Instruct and a stronger external LLM, DeepSeek-V4-Flash, retaining only traces with correct final answers through rejection sampling. As shown in Figure 5(a), these traces improve accuracy from 45.71% to 52.38% and 56.43%, respectively, but remain well below the 63.22% achieved with reasoning derived from heterogeneous experts.

We further find that combining reasoning from five experts increases accuracy from 62.31% with EBM alone to 63.41%, surpassing every individual expert (Figure 5(b)). In contrast, retaining only instances newly solved by each additional expert reaches 62.01%, indicating that the gain cannot be explained by broader sample coverage alone. Together, these results suggest that heterogeneous experts provide diverse inference paths for the same human decision, yielding richer supervision that better reflects the inherent heterogeneity of human social reasoning.

Decision Grounding Matters More Than LLM Generation Freedom. As shown in Figure 5(c), using expert-selected decision evidence achieves 57.28% accuracy, which improves to 58.27% after LLM semantic denoising. Further summarization provides no additional benefit, reaching 57.87%. In contrast, constructing reasoning without expert-selected evidence, delegating the decision to the LLM, or mixing native and expert-guided reasoning reduces accuracy to 55.80%, 54.23%, and 51.52%, respectively. These results suggest that LLMs are most effective as semantic refiners of decision relationships recovered from behavioral data, rather than as unconstrained generators of the underlying decision logic. Detailed settings are provided in Appendix C.2.

A Few Relevant Signals Are Sufficient. Figure 5(d) further shows that effective social reasoning does not require exposing the model to large amounts of information. A single factor selected by feature importance already achieves 62.81% accuracy, while using five factors yields 63.33%. Replacing feature importance with margin-based ranking reduces performance to 62.11%. Together with our earlier observation that more perception is not always better, these results indicate that the key is to identify a small set of decision-relevant signals rather than simply provide more individual information.

Due to space constraints, we provide additional analyses in the Appendix, including response distribution perturbation (Appendix A.3), evidence grounded supervision (Appendix B.5), an ablation without perception information (Appendix A.5), individual experts and repeated exposure (Appendix B.6), reinforcement learning and data selection (Appendix B.7), robustness across model scales (Appendix B.3), and computational efficiency (Appendix B.4).

5 Conclusion

This paper investigates how LLMs can model human decision making and the underlying social reasoning process from structured behavioral data. We introduce **SocialReasoner** and evaluate it across 281,995 respondents and 365 decision questions spanning 13 social domains and more than 40 countries and regions. SocialReasoner improves both individual behavioral fidelity and the recovery of population-level patterns of human behavior, while generalizing to unseen surveys and broader decision domains. These results highlight the potential of structured behavioral data for building foundation models of human social reasoning.

AI USE STATEMENT

We used generative AI tools for language editing, manuscript polishing, and assistance with code development and experimental workflow. LLMs also form part of the experimental methodology, including semantic evidence denoising and reasoning generation. All AI-assisted content, code, analyses, and experimental outputs were reviewed and verified by the authors, who take full responsibility for the final work.

ETHICS STATEMENT

This work studies human social reasoning using existing survey and behavioral data. Because such data may reflect social biases and sensitive individual attributes, models trained on them may reproduce or amplify these patterns. The generated reasoning should be interpreted as a model of human decision processes rather than a verified account of any individual's internal cognition. We discourage applications involving invasive profiling, manipulation, or high-stakes decisions about individuals. All data use and release will follow the licenses, privacy requirements, and access conditions of the original sources.

REPRODUCIBILITY STATEMENT

Details of dataset construction, training, decoding, prompts, and evaluation are provided in the Appendix. Dataset and split details are documented in Appendix E, training and prompt settings in Appendix F.2 and Appendix F.3, and decoding and hardware configurations in Appendix F.1 and Appendix F.4. Code and models will be released to support independent reproduction.

REFERENCES

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*, pp. 337–371. PMLR, 2023.

Diogo Almeida. Introducing system one models & jev. TypeSafe AI Blog, September 2026. URL https://typesafe.ai/blog/introducing-system-one-models-and-jev. Published September 15, 2026.

Anthropic. Introducing claude opus 4.8. https://www.anthropic.com/news/claude-opus-4-8, May 2026. Accessed: 2026-08-16.

Ashwini Ashokkumar, Luke Hewitt, Isaias Ghezae, and Robb Willer. Large language models can predict the results of social science experiments. *Nature*, pp. 1–8, 2026.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025.

Kenneth A Bollen. *Structural equations with latent variables*. John Wiley & Sons, 2014.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3141–3154, 2025.

Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu, Hamid Alinejad-Rokny, Shiwen Ni, et al. Agentcourt: Simulating court with adversarial evolvable lawyer agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5850–5865, 2025.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, et al. Socialbench: Sociality evaluation of role-playing conversational agents. *arXiv preprint arXiv:2403.13679*, 2024.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Michael Davern, Rene Bautista, Jeremy Freese, Pamela Herd, and Stephen L. Morgan. General social survey 1972–2022. NORC ed., Chicago, 2024. 1 datafile (Release 4) and 1 codebook (2022 Release 4). `https://gss.norc.org/content/dam/gss/get-documentation/pdf/codebook/GSS%202022%20Codebook.pdf`.

Robyn M Dawes. The robust beauty of improper linear models in decision making. *American psychologist*, 34(7):571, 1979.

Xuan Long Do, Kenji Kawaguchi, Min-Yen Kan, and Nancy Chen. Aligning large language models with human opinions through persona selection and value–belief–norm reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 2526–2547, 2025.

EVS. European values study 2017: Integrated dataset (evs 2017). GESIS, Cologne, 2022. ZA7500 Data file Version 5.0.0. `https://doi.org/10.4232/1.13897`.

Sara Fish, Yannai A Gonczarowski, and Ran I Shorrer. Algorithmic collusion by large language models. *arXiv preprint arXiv:2404.00806*, 7(2):5, 2024.

Gerd Gigerenzer and Daniel G Goldstein. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4):650, 1996.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in neural information processing systems*, 34:18932–18943, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35: 507–520, 2022.

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

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International conference on artificial intelligence and statistics*, pp. 5549–5581. PMLR, 2023.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the association for computational linguistics: ACL 2023*, pp. 8003–8017, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. Simbench: Benchmarking the ability of large language models to simulate human behaviors. *arXiv preprint arXiv:2510.17516*, 2025.

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

Ji Huang, LI Mengfei, and Shuai Shao. Distribution shift alignment helps llms simulate survey response distributions. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9395–9409, 2026.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5906–5919, 2023.

ISSP Research Group. International social survey programme: Social inequality iv - issp 2009. GESIS Data Archive, Cologne, 2017a. ZA5400 Data file Version 4.0.0. https://doi.org/10.4232/1.12777.

ISSP Research Group. International social survey programme: Work orientations iv - issp 2015. GESIS Data Archive, Cologne, 2017b. ZA6770 Data file Version 2.1.0. https://doi.org/10.4232/1.12848.

ISSP Research Group. International social survey programme: Role of government v - issp 2016. GESIS Data Archive, Cologne, 2018. ZA6900 Data file Version 2.0.0. https://doi.org/10.4232/1.13052.

ISSP Research Group. International social survey programme: Social inequality v - issp 2019. GESIS, Cologne, 2022. ZA7600 Data file Version 3.0.0. https://doi.org/10.4232/1.14009.

ISSP Research Group. International social survey programme: Health and health care ii - issp 2021. GESIS, Cologne, 2024. ZA8000 Data file Version 2.0.0. https://doi.org/10.4232/5.ZA8000.2.0.0.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S Bernstein. Finetuning llms for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30084–30099, 2025.

Rishikesh Konapure. Home loan approval. Kaggle Dataset, 2023. Accessed: 2026-08-30.

Taeyoon Kwon, Kai Tzu-iunn Ong, Dongjin Kang, Seungjun Moon, Jeong Ryong Lee, Dosik Hwang, Beomseok Sohn, Yongsik Sim, Dongha Lee, and Jinyoung Yeo. Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 18417–18425, 2024.

Chance Jiajie Li, Jiayi Wu, Zhenze Mo, Ao Qu, Yuhan Tang, Kaiya Zhao, Yulu Gan, Jie Fan, Jiangbo Yu, Jinhua Zhao, et al. Simulating society requires simulating thought. *Advances in Neural Information Processing Systems*, 38, 2026a.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024a.

Ming Li, Xirui Li, and Tianyi Zhou. Does socialization emerge in ai agent society? a case study of moltbook. In *Proceedings of the ACM Conference on AI and Agentic Systems*, pp. 537–550, 2026b.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. Econagent: large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 15523–15536, 2024b.

Tianyuan Liang, Xuwei Tan, Lei Shi, Junsheng Zhong, Ziyu Hu, Tian Xie, Zhiqun Zuo, Xiaodong Yu, and Xueru Zhang. Tlrd: Teaching llms to reason over tabular data with tri-level rationale distillation. *arXiv preprint arXiv:2606.08295*, 2026.

Chenxi Lin, Zhuoren Jiang, Kaisong Song, and Yiquan Wu. Safo: Stable adaptive fairness optimization for llm-based social survey simulation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31626–31654, 2026.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025a.

Ryan Liu, Jiayi Geng, Joshua C Peterson, Ilia Sucholutsky, and Thomas L Griffiths. Large language models assume people are more rational than we really are. *arXiv preprint arXiv:2406.17055*, 2024.

Zhaowei Liu, Xin Guo, Zhi Yang, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Mengping Li, Qi Qi, Zhiqiang Liu, Yiyang Han, et al. Fin-r1: A large language model for financial reasoning through reinforcement learning. *arXiv preprint arXiv:2503.16252*, 2025b.

Alejandro Lopez-Lira and Yuehua Tang. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*, 2023.

Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O Jackson. A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9): e2313925121, 2024.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.

Robert M Nosofsky. Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, 115(1):39, 1986.

Keane Ong, Sabri Boughorbel, Luwei Xiao, Chanakya Ekbote, Wei Dai, Ao Qu, Jingyao Wu, Rui Mao, Ehsan Hoque, Erik Cambria, et al. Omnisapiens: A foundation model for social behavior processing via heterogeneity-aware relative policy optimization. *arXiv preprint arXiv:2602.10635*, 2026.

OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`, July 2024.

OpenAI. Gpt-5.6: Frontier intelligence that scales with your ambition. `https://openai.com/index/gpt-5-6/`, July 2026a.

OpenAI. GPT-6 Astra System Card, September 2026b. URL `https://deploymentsafety.openai.com/gpt-6-astra`.

John W Payne, James R Bettman, and Eric J Johnson. Adaptive strategy selection in decision making. *Journal of experimental psychology: Learning, Memory, and Cognition*, 14(3):534, 1988.

Pew Research Center. American trends panel wave 142. `https://www.pewresearch.org/dataset/american-trends-panel-wave-142/`, 2024a. Fielded February 7–11, 2024.

Pew Research Center. American trends panel wave 152. https://www.pewresearch.org/dataset/american-trends-panel-wave-152/, 2024b. Fielded August 12–18, 2024.

Jinghua Piao, Cheng Huang, Jiarui Huang, Songwei Li, Ranran Wang, Yong Li, et al. Marketsim: Simulating stock markets with large-scale generative agents. In *Forty-third International Conference on Machine Learning*.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025.

Jinghua Piao, Jun Zhang, Haoyu Huang, Keming Zhang, Jing Yi Wang, Songwei Li, Boyuan Sun, Jiayi Chang, Fengli Xu, Chunyan Wang, et al. Agentsociety 2: An integrated research environment for executable social science. *arXiv preprint arXiv:2607.11895*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International conference on machine learning*, pp. 29971–30004. PMLR, 2023.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21147–21170, 2025.

Joseph Suh, Suhong Moon, and Serina Chang. Graph-based alternatives to llms for human simulation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46479–46506, 2026.

Ala N Tak, Amin Banayeeanzade, Anahita Bolourani, Fatemeh Bahrani, Ashutosh Chaubey, Sai Praneeth Karimireddy, Norbert Schwarz, and Jonathan Gratch. Sparks of rationality: Do reasoning llms align with human judgment and choice? *arXiv preprint arXiv:2601.22329*, 2026.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Qwen Team. Qwen3. 6-27b: Flagship-level coding in a 27b dense model, 2026.

Qwen Team et al. Qwq-32b: Embracing the power of reinforcement learning, 2025.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411, 2025a.

Jia Wang, Ziyu Zhao, Tingjuntao Ni, and Zhongyu Wei. Sociobench: Modeling human behavior in sociological surveys with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26268–26300, 2025b.

Lei Wang, Heyang Gao, Xiaohe Bo, Xu Chen, and Ji-Rong Wen. Yulan-onesim: Towards the next generation of social simulator with large language models. In *Workshop on Scaling Environments for Agents*, 2025c.

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.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. Finben: A holistic financial benchmark for large language models. *Advances in neural information processing systems*, 37:95716–95743, 2024.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan Suchow, Zhenyu Cui, Rong Liu, et al. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37:137010–137045, 2024.

Ziyun Yu, Yiru Zhou, Chen Zhao, and Hongyi Wen. An analysis of large language models for simulating user responses in surveys. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 242–259, 2025.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. On generative agents in recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval*, pp. 1807–1817, 2024a.

Linghao Zhang, Shilin He, Chaoyun Zhang, Yu Kang, Bowen Li, Chengxing Xie, Junhao Wang, Maoquan Wang, Yufan Huang, Shengyu Fu, et al. Swe-bench goes live! *Advances in Neural Information Processing Systems*, 38, 2026.

Xinnong Zhang, Jiayu Lin, Libo Sun, Weihong Qi, Yihang Yang, Yue Chen, Hanjia Lyu, Xinyi Mou, Siming Chen, Jiebo Luo, et al. Electionsim: Massive population election simulation powered by large language model driven agents. *arXiv preprint arXiv:2410.20746*, 2024b.

Xuanyu Zhang, Qing Yang, and Dongliang Xu. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. *arXiv preprint arXiv:2305.12002*, 2023.

Zihan Zheng, Zerui Cheng, Zeyu Shen, Shang Zhou, Kaiyuan Liu, Hansen He, Dongruixuan Li, Stanley Wei, Hangyi Hao, Jianzhu Yao, et al. Livecodebench pro: How do olympiad medalists judge llms in competitive programming? *arXiv preprint arXiv:2506.11928*, 2025.

Xuhui Zhou, Zhe Su, Sophie Feng, Jiaxu Zhou, Jen-tse Huang, Hsien-Te Kao, Spencer Lynch, Svitlana Volkova, Tongshuang Wu, Anita Woolley, et al. Sotopia-s4: a user-friendly system for flexible, customizable, and large-scale social simulation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 350–360, 2025.

Jie Zhu, Qian Chen, Huaixia Dou, Junhui Li, Lifan Guo, Feng Chen, and Chi Zhang. Dianjin-r1: Evaluating and enhancing financial reasoning in large language models. *arXiv preprint arXiv:2504.15716*, 2025.

Caleb Ziems, William Held, Su Doga Karaca, David Grusky, Tatsunori Hashimoto, and Diyi Yang. Will scaling improve social simulation with llms? *arXiv preprint arXiv:2607.02464*, 2026.

CONTENTS

A Supplementary Motivation and Evidence **18**

A.1 Cognitive Baselines for Survey Response Prediction 18

A.2 Why Perception Information Matters for Social Reasoning 20

A.3 Does SFT Rely on Response Distribution Regularities? 21

A.4 Revisiting Subpopulation Modeling: The Role of Cross-Group Response Distributions 22

A.5 The Limits of Demographic-Only Social Reasoning 24

A.6 Performance of Reasoning Models on Social Reasoning Benchmarks 25

A.7 Does Social Reasoning Benefit from Model Scaling? 26

B Additional Experimental Results and Ablations **28**

B.1 Population-Level Survey Response Distribution Prediction 28

B.2 Detailed Analysis of Machine Learning Models 30

B.3 Robustness Across Model Scales 34

B.4 Efficiency Analysis 35

B.5 Effect of Evidence-Grounded Supervision 36

B.6 Individual Expert Supervision and Repeated Exposure 37

B.7 Effect of Reinforcement Learning and Data Selection 38

C Method Details **39**

C.1 Feature Attribution for Decision Evidence Construction 39

C.2 The Role of LLM Intervention in Social Reasoning Construction 42

D Detailed Discussion of Related Works **46**

E Datasets Details **48**

E.1 American Trends Panel 48

E.2 European Values Study 49

E.3 International Social Survey Programme 49

E.4 General Social Survey 51

F Details On Experimental Settings **51**

F.1 Decoding and Sampling Settings 51

F.2 Training Hyperparameters 51

F.3 Prompt Details 51

F.4 Hardware Configuration 53

G Case Study **54**

A Supplementary Motivation and Evidence

A.1 Cognitive Baselines for Survey Response Prediction

Summary. We evaluate four classical cognitive models as question specific baselines for survey response prediction. These models capture simple forms of cue weighting, cue selection, and exemplar similarity, providing interpretable references for prediction from structured respondent information. Among them, Take-the-Best is particularly competitive with several LLM baselines. We interpret these results as evidence that simple decision rules can capture substantial predictive structure in survey responses, rather than as evidence that respondents necessarily follow these cognitive mechanisms.

We consider the Weighted Additive Rule (WADD) (Payne et al., 1988), Equal Weight Rule (EW) (Dawes, 1979), Take-the-Best (TTB) (Gigerenzer & Goldstein, 1996), and the Generalized Context Model (GCM) (Nosofsky, 1986). WADD, EW, and TTB were originally formulated for binary decisions. Because survey questions may contain more than two response options, we adapt them through pairwise comparisons and aggregate the resulting decisions. GCM naturally supports multiple response options and is applied directly as an exemplar model.

For a given survey question, let $\mathcal{Y}$ denote its response options. Each respondent i is represented by a profile

$$\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iD}), \tag{6}$$

where each z_{ij} is a demographic or perception attribute. All statistics are estimated separately for each target question using only the training respondents for that question.

Pairwise Cue Construction. WADD, EW, and TTB use the same pairwise cue representation. Consider two response options $a, b \in \mathcal{Y}$. For cue j and one of its categorical values v, we count the training respondents with that value who selected each option:

$$N^a_{j,v} = \sum_{i=1}^{N} \mathbf{1}[y_i = a]\,\mathbf{1}[z_{ij} = v]\,, \qquad N^b_{j,v} = \sum_{i=1}^{N} \mathbf{1}[y_i = b]\,\mathbf{1}[z_{ij} = v]\,. \tag{7}$$

The direction of cue value v is determined by which response option is more frequent among respondents with that value:

$$d_j^{(a,b)}(v) = \begin{cases} +1, & N^a_{j,v} > N^b_{j,v}, \\ -1, & N^a_{j,v} < N^b_{j,v}, \\ 0, & N^a_{j,v} = N^b_{j,v}. \end{cases} \tag{8}$$

A positive direction supports option a, a negative direction supports option b, and a zero direction indicates that the cue value does not distinguish between the two options.

We measure the reliability of cue j using its pair specific validity:

$$\nu_j^{(a,b)} = \frac{\sum\limits_{v:\, d_j^{(a,b)}(v)\neq 0} \max\{N^a_{j,v}, N^b_{j,v}\}}{\sum\limits_{v:\, d_j^{(a,b)}(v)\neq 0} \left(N^a_{j,v} + N^b_{j,v}\right)}. \tag{9}$$

This quantity corresponds to the empirical accuracy obtained by assigning each discriminating cue value to its majority response option. Cue values with equal counts are excluded because they provide no directional evidence. If a cue never distinguishes a from b, its validity is set to zero. Missing and previously unseen cue values also contribute zero directional evidence.

Weighted Additive Rule. WADD combines evidence from all available cues and weights each cue by its pair specific validity:

$$S^{(a,b)}_{\mathrm{WADD}}(\mathbf{z}_i) = \sum_{j=1}^{D} \nu_j^{(a,b)} d_j^{(a,b)}(z_{ij}). \tag{10}$$

A positive score selects option a, whereas a negative score selects option b. Cues that more reliably distinguish the two options therefore contribute more strongly to the pairwise prediction.

Equal Weight Rule. EW uses the same cue directions but assigns equal weight to every cue:

$$S^{(a,b)}_{\mathrm{EW}}(\mathbf{z}_i) = \sum_{j=1}^{D} d_j^{(a,b)}(z_{ij}). \tag{11}$$

As in WADD, a positive score selects a, whereas a negative score selects b. Comparing WADD with EW therefore measures whether weighting cues by their empirical validity improves prediction over equally weighted directional evidence.

Take-the-Best. TTB follows a one reason decision strategy. For each option pair (a, b), cues are ranked from highest to lowest pair specific validity:

$$\nu_{\pi_1}^{(a,b)} \geq \nu_{\pi_2}^{(a,b)} \geq \cdots \geq \nu_{\pi_D}^{(a,b)}. \tag{12}$$

For respondent i, TTB examines cues in this order and selects the first cue satisfying

$$d_{\pi_t}^{(a,b)}(z_{i,\pi_t}) \neq 0.$$

The sign of this cue determines the pairwise prediction. A positive direction selects a, whereas a negative direction selects b. Once a discriminating cue is found, lower ranked cues are ignored. If no cue distinguishes the two options, the comparison results in a tie.

Multiclass Aggregation. WADD, EW, and TTB produce one binary decision for each pair of response options. To obtain a prediction for a multiclass survey question, we compare all option pairs and aggregate the results as a round robin tournament.

For model $m \in \{\mathrm{WADD}, \mathrm{EW}, \mathrm{TTB}\}$, let $h_m^{(k,\ell)}(\mathbf{z}_i)$ denote the winner between options k and ℓ. The number of pairwise wins received by option k is

$$V_{i,k}^{(m)} = \sum_{\ell \in \mathcal{Y} \setminus \{k\}} \mathbf{1}\left[h_m^{(k,\ell)}(\mathbf{z}_i) = k\right]. \tag{13}$$

The final prediction is the option with the largest number of pairwise wins:

$$\widehat{y}_i^{(m)} = \arg\max_{k \in \mathcal{Y}} V_{i,k}^{(m)}. \tag{14}$$

Pairwise and final tournament ties are resolved using a fixed seeded procedure. No class prior or additional confidence heuristic is used.

Equal Weight Generalized Context Model. GCM uses exemplar similarity rather than cue directions. A target respondent is compared with training respondents, and more similar profiles contribute more strongly to the predicted response.

We use an equal weight GCM so that all respondent attributes contribute equally to the distance calculation. For a target respondent $\mathbf{z}$ and a training respondent $\mathbf{z}_r$, let $\mathcal{O}(\mathbf{z}, \mathbf{z}_r)$ denote the attributes observed for both respondents. Their categorical distance is

$$d(\mathbf{z}, \mathbf{z}_r) = \frac{\sum_{j \in \mathcal{O}(\mathbf{z}, \mathbf{z}_r)} \mathbf{1}[z_j \neq z_{rj}]}{|\mathcal{O}(\mathbf{z}, \mathbf{z}_r)|}. \tag{15}$$

The distance is the proportion of jointly observed attributes on which the two respondents differ. If no attribute is jointly observed, the distance is set to one.

We convert distance into similarity using exponential decay:

$$s(\mathbf{z}, \mathbf{z}_r) = \exp[-c\, d(\mathbf{z}, \mathbf{z}_r)]\,, \tag{16}$$

where $c > 0$ controls the rate of decay. We fix $c = 5$ across all questions and datasets to avoid question specific tuning.

For response option k, GCM accumulates similarity to training respondents who selected k:

$$E_k(\mathbf{z}) = \sum_{r:\, y_r = k} s(\mathbf{z}, \mathbf{z}_r). \tag{17}$$

The predicted response is the option with the greatest accumulated exemplar evidence:

$$\widehat{y}_{\mathrm{GCM}} = \arg\max_{k \in \mathcal{Y}} E_k(\mathbf{z}). \tag{18}$$

Discussion. As shown in Table 1, these simple cognitive baselines provide nontrivial performance on survey response prediction. TTB is particularly competitive, outperforming several LLM-based baselines for human social reasoning on multiple datasets. This result indicates that a small number of highly predictive cues can capture substantial structure in survey responses.

At the same time, these cognitive baselines remain below the strongest question specific machine learning models analyzed in Appendix B.2. This gap indicates that the simple decision rules considered here do not capture all predictive structure in the data. More flexible models can represent relationships among demographic and perception attributes that are not captured by independent cue weighting, single cue selection, or equal weight exemplar similarity.

Overall, the comparison provides a useful reference point for LLM based human social reasoning. Strong general reasoning capability does not automatically translate into stronger individual response prediction than simple models built directly from structured behavioral data. This observation motivates our use of machine learning experts to extract predictive evidence from survey data before transferring it to LLMs.

A.2 Why Perception Information Matters for Social Reasoning

Summary. Perception information provides substantial predictive signal for individual response prediction beyond demographic attributes. Across five evaluated LLMs, removing perception information reduces average accuracy by 10.69 percentage points. Structural equation modeling further shows that perception variables are more strongly associated with observed responses than demographic attributes. Together, these results motivate preserving perception information when constructing datasets for human social reasoning.

Table 3 evaluates LLMs using demographic information without perception information. Removing perceptions causes substantial performance degradation across all five evaluated models, with an average accuracy drop of 10.69 percentage points. This result shows that demographic attributes alone contain substantially less information for predicting individual responses in our setting.

This finding is consistent with the structural analysis in Figure 2(b). The SEM results show that perception variables are more strongly associated with observed responses, whereas the direct associations of demographic attributes are generally weaker. These two analyses provide complementary evidence. The input ablation measures the predictive contribution of perception information to

Table 3: **Comparison of LLM performance with and without human perception information.** We report the complete results under both input settings. Δ denotes the accuracy-point change after removing human perception information, computed as *Without Perception* minus *With Perception*. Red values indicate performance degradation.

Input Setting	ATP142	ATP152	EVS2017	ISSP2009	ISSP2015	ISSP2016	ISSP2019	ISSP2021	Avg.
GPT-4o-mini									
With Perception	51.53	54.38	51.15	46.11	44.82	43.62	39.16	40.62	46.42
Without Perception	40.34	31.97	47.39	32.05	31.48	33.81	30.95	27.95	34.49
Δ	-11.19	-22.41	-3.76	-14.06	-13.34	-9.81	-8.21	-12.67	-11.93
DeepSeek-V3.2									
With Perception	50.65	54.04	52.88	46.70	38.26	43.01	39.81	36.92	45.28
Without Perception	41.64	33.93	49.49	30.28	33.12	31.89	29.77	29.42	34.94
Δ	-9.01	-20.11	-3.39	-16.42	-5.14	-11.12	-10.04	-7.50	-10.34
Llama-3.1-70B-Instruct									
With Perception	51.21	52.33	55.03	46.54	42.94	44.63	41.83	42.39	47.11
Without Perception	43.21	34.58	43.75	32.76	29.70	32.94	31.62	30.57	34.89
Δ	-8.00	-17.75	-11.28	-13.78	-13.24	-11.69	-10.21	-11.82	-12.22
Qwen3-4B-2507-Instruct									
With Perception	47.51	49.81	44.88	44.68	39.06	41.56	40.46	38.95	43.36
Without Perception	37.27	27.01	40.71	30.03	32.66	28.25	30.51	29.86	32.04
Δ	-10.24	-22.80	-4.17	-14.65	-6.40	-13.31	-9.95	-9.09	-11.33
Qwen2.5-7B-Instruct									
With Perception	47.30	45.71	44.00	42.95	33.38	35.77	39.05	33.29	40.18
Without Perception	38.59	28.56	40.26	33.68	29.57	30.31	31.61	27.64	32.53
Δ	-8.71	-17.15	-3.74	-9.27	-3.81	-5.46	-7.44	-5.65	-7.65

LLM-based modeling of human social reasoning, while the SEM analysis characterizes its statistical relationship with human responses.

Recent work has argued that modeling human behavior requires going beyond static demographic attributes to capture internal cognitive states (Li et al., 2026a). Our results provide direct empirical support for this view in individual response prediction. Individuals with similar demographic backgrounds can still differ substantially in their beliefs, attitudes, evaluations, concerns, and expectations. Demographic attributes therefore provide useful background information but do not fully distinguish how individuals respond to a particular issue. Perception measurements provide additional respondent specific information that helps capture this variation.

This distinction is also relevant to LLM based social simulation, where individuals are often represented through profile descriptions containing demographic, role, or background information (Piao et al., 2025; 2026; Li et al., 2026b; Wang et al., 2025c). Our results suggest that enriching such representations with observed perceptions can provide additional information for modeling heterogeneous individual responses, rather than relying on demographic characteristics alone.

A.3 Does SFT Rely on Response Distribution Regularities?

Summary. We examine whether the gains of standard SFT are sensitive to response distribution regularities in the training data. After balancing response option frequencies, the improvement of SFT over the base model drops from 9.08 to -1.30 percentage points, whereas SocialReasoner retains 11.21 points of its original 13.29-point gain. This contrast indicates that SocialReasoner is substantially less dependent on option frequency patterns for individual response prediction.

To test the effect of response distribution regularities, we rebalance the ATP152 training set by equalizing the frequencies of response options while leaving the test set unchanged. This intervention reduces option frequency patterns in the training data without altering the evaluation distribution. As shown in Table 4, standard SFT improves over the base model by 9.08 percentage points under the original training distribution, but falls to 1.30 points below the base model after rebalancing. In contrast, SocialReasoner retains an 11.21-point improvement after rebalancing, compared with 13.29 points under the original distribution.

The sharp degradation of SFT shows that its in-domain performance is highly sensitive to response regularities in the training data. SocialReasoner is considerably more stable under the same inter-

Table 4: **Effect of balancing response-option frequencies in the training data.** All methods are evaluated on the same unchanged ATP152 test set. Balancing the training response distribution removes option-frequency shortcuts while preserving the evaluation distribution.

Method	Original Train	Balanced Train	Gain over Base
Qwen3-4B	49.81	49.81	–
SFT	58.89	48.51	**+9.08 → -1.30**
Ours	**63.10**	**61.02**	**+13.29 → +11.21**

vention, preserving most of its original improvement when option frequency patterns are reduced. These results are consistent with SocialReasoner relying less on answer distribution regularities and more on respondent information associated with individual choices.

A.4 Revisiting Subpopulation Modeling: The Role of Cross-Group Response Distributions

Summary. We examine why conventional subpopulation modeling remains weaker than supervised fine tuning (SFT) for individual response prediction. By conditioning response distributions on combinations of respondent attributes, a simple cross-group estimator recovers a substantial portion of the in-domain advantage of SFT. This result shows that fine-grained conditional response regularities alone can support strong individual prediction, and that high prediction accuracy does not by itself imply that intermediate decision evidence has been explicitly modeled.

In the preceding analysis, conventional subpopulation prediction remains noticeably weaker than SFT in individual response accuracy. We therefore examine whether this gap is related to the granularity at which population variation is represented. Standard SubPop methods typically estimate response distributions for one respondent group at a time. Such distributions capture broad population differences but can obscure variation that emerges only when multiple respondent characteristics are considered jointly.

From Single-Group to Cross-Group Distributions. For a target question q, consider a respondent attribute a taking value v. A conventional subpopulation estimator models the empirical response distribution

$$\widehat{P}(y \mid q, a = v), \tag{19}$$

and uses it as a group level prior for respondents satisfying $a = v$. Because the remaining respondent characteristics are marginalized out, the same distribution can contain respondents with substantially different response tendencies.

Let z denote additional respondent characteristics. Then

$$\widehat{P}(y \mid q, a = v) = \sum_{z} \widehat{P}(y \mid q, a = v, z)\widehat{P}(z \mid q, a = v). \tag{20}$$

When response distributions differ across combinations of attributes, this marginalization removes information that may be useful for individual prediction. This motivates estimating finer conditional distributions over intersections of respondent characteristics.

Hierarchical Cross-Group Population Modeling. To examine this effect, we construct a hierarchical population library that retains combinations of respondent attributes when they correspond to a sufficiently large change in the empirical response distribution.

For each target question, a population group is represented as

$$g = \{(a_1 = v_1), \ldots, (a_k = v_k)\}, \tag{21}$$

with empirical response distribution

$$\widehat{P}_g(y \mid q) = \frac{N_{q,g,y}}{\sum_{y'} N_{q,g,y'}}, \tag{22}$$

where $N_{q,g,y}$ denotes the number of respondents satisfying all conditions in g who select response y for question q.

We first construct all one attribute groups with sufficient sample support. Higher order groups are then generated recursively from retained lower order groups. For a candidate group g and one of its retained parent groups g', we measure the change in response distribution using total variation distance:

$$\Delta(g, g') = \frac{1}{2} \sum_{y} \left| \widehat{P}_g(y \mid q) - \widehat{P}_{g'}(y \mid q) \right|. \tag{23}$$

A candidate group is retained only when it satisfies both a minimum support constraint and a minimum distributional shift:

$$N_{q,g} \geq n_{\min}, \qquad \max_{g' \in \mathrm{Par}(g)} \Delta(g, g') \geq \tau. \tag{24}$$

We use $n_{\min} = 100$ and $\tau = 0.15$. Only retained groups are expanded further, and construction stops when no higher order group satisfies the retention criterion. This procedure preserves attribute intersections associated with meaningful changes in response distributions while avoiding exhaustive enumeration of the full combinatorial space.

Fine-Grained Response Distributions Recover Much of the SFT Gain. Cross-group population modeling substantially reduces the individual prediction gap between conventional SubPop and SFT. Notably, the cross-group estimator achieves this improvement by conditioning empirical response distributions on increasingly specific combinations of observed respondent characteristics rather than by learning an explicit reasoning process.

This comparison provides a useful perspective on the strong in-domain performance of SFT. A substantial portion of its advantage can be reproduced by a lightweight estimator that captures sufficiently fine-grained conditional response regularities. The result does not imply that SFT explicitly implements the same population estimator, nor that it memorizes individual training examples. Instead, it shows that detailed relationships between respondent characteristics and response frequencies are themselves sufficient to produce strong individual prediction within the observed survey.

Consequently, high in-domain accuracy alone does not distinguish between models that exploit detailed conditional response regularities and models that explicitly represent evidence connecting individual information to a choice. Additional analysis is therefore needed to understand what form of supervision contributes beyond answer level prediction.

Implications for Evidence Supervision. This observation motivates an important distinction between answer supervision and evidence supervision. Both fine-grained population estimators and standard SFT are trained from observed responses and can benefit from regularities in the source data. They do not, however, explicitly supervise which respondent information supports a particular prediction.

SocialReasoner introduces this additional supervision. Rather than providing only the final response, it identifies predictive evidence from structured behavioral data and converts that evidence into natural language reasoning traces. The student therefore receives supervision not only on which response is observed, but also on which information is predictive of that response.

The comparison reveals a progression in the information provided to the model. Conventional SubPop captures relatively coarse response distributions, while cross-group modeling captures finer

conditional regularities and recovers much of the in-domain advantage of SFT. SocialReasoner further exposes predictive evidence associated with individual responses, providing a richer supervision signal than the final answer alone.

A.5 The Limits of Demographic-Only Social Reasoning

Summary. SocialReasoner remains effective when only demographic information is available, showing that the framework does not depend on perception information to extract useful predictive evidence. However, demographic attributes provide a much coarser basis for individual prediction. Qualitative examples further show that reasoning constructed from demographic evidence can turn statistical associations into plausible individual explanations. We therefore interpret demographic-only reasoning as predictive supervision rather than evidence of the underlying cognitive process of a respondent.

SocialReasoner is designed to transfer predictive knowledge learned from structured behavioral data into LLMs and does not require perception information as a necessary input. We therefore evaluate the same framework using only demographic attributes. As shown in Figure 6, SocialReasoner consistently improves over the base model and the comparison methods across all eight datasets. This result shows that the proposed knowledge transfer mechanism remains effective even when the available respondent information is restricted to demographics.

At the same time, demographic-only prediction has an important limitation. Attributes such as age, gender, race or ethnicity, education, income, and residential context can be predictive of survey responses at the population level, but these associations do not uniquely determine how a particular individual evaluates a specific issue. When such associations are converted into natural language reasoning, they can produce coherent explanations that go beyond what is directly supported by the observed information. The following examples illustrate this behavior.

Illustrative Demographic-Only Rationales

Example 1

Question: Thinking about the use of artificial intelligence (AI) in the United States, are you more concerned that the U.S. government will...

Options:
(1) Go too far regulating its use
(2) Not go far enough regulating its use
(3) Not sure

Ground Truth Answer: Not go far enough regulating its use.

Generated CoT:
As a Hispanic male with some college education living in a suburban area of the Western United States, I am generally aware of the potential benefits and risks associated with AI. Given my moderate political ideology and Republican-leaning views, I tend to believe that there should be some regulation but also recognize the importance of not stifling innovation. Therefore, I am more concerned that the government might not regulate AI sufficiently to protect privacy and prevent misuse.

Example 2

Question: When it comes to artificial intelligence (AI), how concerned are you about AI leading to less connection between people?

Options:
(1) Extremely concerned
(2) Very concerned
(3) Somewhat concerned
(4) Not too concerned
(5) Not at all concerned

Ground Truth Answer: Somewhat concerned.

Generated CoT:
As a moderately concerned individual with some college education and a background in a rural area, I recognize the potential impacts of AI on social interactions, but I do not see it as an extreme threat. My Protestant faith and infrequent church attendance suggest a more practical approach to societal changes rather than an overly fearful one. Additionally, my lack of volunteer activities and lower income position may indicate a focus on immediate needs over broader concerns about technology's effects on society.

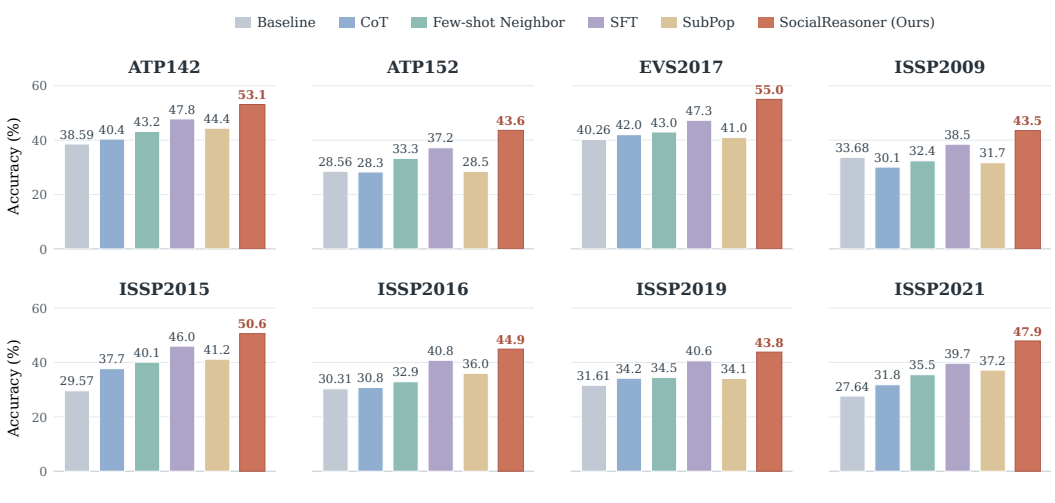

Figure 6: **Performance without perception information.** We evaluate Qwen2.5-7B-Instruct using only demographic information across eight survey datasets. SocialReasoner consistently improves over the base model and the comparison methods, showing that the framework remains effective when perception information is unavailable.

These examples should not be interpreted as evidence that the demographic attributes shown above causally produce the corresponding beliefs or choices. Rather, they illustrate a general risk of demographic-only reasoning: predictive associations present in behavioral data can be verbalized as individual-level explanations even when the observed attributes do not directly specify those explanations. The resulting rationale may therefore be useful for prediction while still exceeding what can be justified from the respondent information alone.

Perception information reduces this ambiguity by providing observations that are more closely related to the issue being predicted. Respondents with similar demographic backgrounds may differ substantially in their beliefs, attitudes, concerns, expectations, and evaluations. As shown in Section A.2, these perception variables provide substantial additional predictive signal beyond demographic attributes. They therefore offer a richer basis for constructing evidence that distinguishes individual responses.

Taken together, these results separate two aspects of SocialReasoner. The framework can extract and transfer useful predictive information even from demographic attributes alone, demonstrating that its effectiveness is not contingent on perception data. However, the quality of the resulting evidence depends on the information available in the input. Perception information provides a more informative basis for individual response prediction and reduces the need to rely primarily on broad demographic regularities.

A.6 Performance of Reasoning Models on Social Reasoning Benchmarks

Summary. Long reasoning does not consistently improve human social reasoning performance. Across several representative reasoning model families, substantially longer outputs do not translate into consistently higher response prediction accuracy, despite much greater inference cost. These results motivate our focus on the content of reasoning supervision rather than reasoning length: SocialReasoner grounds concise reasoning traces in predictive evidence extracted from behavioral data.

We evaluate a broad set of representative reasoning models on our human social reasoning benchmarks, including QwQ-32B (Team et al., 2025), the DeepSeek-R1 family (Guo et al., 2025), the Qwen3 family (Yang et al., 2025), and the Qwen3.5 family (Qwen Team, 2026). Standard models and SocialReasoner use a maximum generation length of 512 tokens. For the reasoning models evaluated in this section, we increase this limit to 10,000 tokens because a 512-token budget frequently truncates their reasoning before a final response is produced. All other decoding settings

Table 5: **Performance of reasoning models on social reasoning benchmarks.** We report individual response prediction accuracy (%) across eight perception-aware survey datasets.

Model	ATP142	ATP152	EVS2017	ISSP2009	ISSP2015	ISSP2016	ISSP2019	ISSP2021	Avg.
DeepSeek-R1 Distill Family									
DeepSeek-R1-Distill-Qwen-1.5B	28.74	27.54	28.63	24.51	23.87	27.32	24.75	25.02	26.30
DeepSeek-R1-Distill-Qwen-7B	38.06	39.66	41.38	35.33	26.23	30.90	31.74	28.89	34.02
DeepSeek-R1-Distill-Llama-8B	45.46	45.27	41.59	40.72	28.45	38.78	34.57	33.09	38.49
DeepSeek-R1-Distill-Qwen-14B	48.39	53.98	48.62	43.23	38.30	41.42	37.92	40.27	44.02
QwQ Family									
QwQ-32B	47.17	49.77	54.41	45.88	37.42	41.62	38.17	36.23	43.83
Qwen3 Family									
Qwen3-0.6B	27.31	33.77	30.01	18.51	18.07	18.68	20.80	17.34	23.06
Qwen3-4B-Thinking-2507	43.52	55.22	50.08	44.92	37.10	39.55	38.57	35.84	43.10
Qwen3-8B	43.15	53.15	48.65	48.45	38.36	42.11	40.33	39.37	44.20
Qwen3-14B	44.89	48.24	50.52	47.50	38.49	42.68	39.63	37.73	43.71
Qwen3.5 Family									
Qwen3.5-0.8B	37.35	26.50	31.34	26.13	24.22	24.71	24.69	24.57	27.44
Qwen3.5-2B	41.66	43.43	40.22	36.48	34.78	33.13	29.56	30.82	36.26
Qwen3.5-4B	46.15	46.93	44.57	41.72	35.11	37.00	35.37	31.55	39.80
Qwen3.5-9B	39.17	52.67	46.49	39.18	33.98	35.33	33.20	31.33	38.92

remain unchanged. The comparison therefore evaluates these models with sufficient room for their characteristic long reasoning behavior.

As shown in Table 5, reasoning models do not consistently achieve higher individual response prediction accuracy than standard instruction tuned models. The difference in generation cost, however, is substantial. Standard instruction tuned models typically produce only about 2 output tokens when directly predicting a response, explicit CoT prompting produces approximately 384 tokens on average, and long reasoning models commonly generate around 4,000–5,000 tokens per response. We do not observe a consistent relationship between this increase in reasoning length and prediction accuracy. Thus, generating substantially longer reasoning traces is not by itself sufficient to improve human social reasoning performance.

This distinction is important for large scale social simulation, where a single experiment may require predictions for many individuals or repeated interactions over multiple rounds. Long reasoning traces substantially increase inference cost and latency, while the results in Table 5 show no corresponding consistent gain in response prediction accuracy.

These results do not conflict with our use of reasoning supervision in SocialReasoner. The relevant distinction is between reasoning length and reasoning content. Generic reasoning models generate extended deliberation using knowledge acquired during pretraining and post training. SocialReasoner instead constructs reasoning supervision from predictive relationships learned directly from structured behavioral data. The goal is therefore not to encourage more reasoning, but to make information associated with the observed response explicit in the supervision.

Specifically, SocialReasoner first identifies predictive evidence using machine learning experts and then converts the retained evidence into concise reasoning traces. This produces substantially shorter supervision than the reasoning generated by long reasoning models while directly grounding the trace in respondent information associated with the target choice. Together with the results above, this suggests that human social reasoning benefits more from decision relevant evidence than from increasing reasoning length alone.

A.7 Does Social Reasoning Benefit from Model Scaling?

Summary. Model scale is strongly associated with response prediction accuracy on human social reasoning tasks across the full model range, but this relationship weakens substantially among larger models. Across 32 open weight models, the correlation between total parameter count and average accuracy is $\rho = 0.896$, while for models with more than 32B parameters it decreases to $\rho = 0.300$. These results suggest that larger model capacity alone does not provide a consistent advantage for

Table 6: **Scaling behavior of open-weight models on social reasoning benchmarks.** We report individual response prediction accuracy (%) across eight perception-aware survey datasets. Models are ordered by total parameter count. For MoE models, parameter counts are reported as total/activated parameters.

Model	Params (B)	ATP142	ATP152	EVS2017	ISSP2009	ISSP2015	ISSP2016	ISSP2019	ISSP2021	Avg.
Qwen3-0.6B	0.6	27.31	33.77	30.01	18.51	18.07	18.68	20.80	17.34	23.06
Qwen3.5-0.8B	0.8	37.35	26.50	31.34	26.13	24.22	24.71	24.69	24.57	27.44
DeepSeek-R1-Distill-Qwen-1.5B	1.5	28.74	27.54	28.63	24.51	23.87	27.32	24.75	25.02	26.30
Qwen2.5-1.5B-Instruct	1.5	28.94	25.99	27.40	21.73	19.52	20.57	21.84	22.22	23.53
Qwen3.5-2B	2	41.66	43.43	40.22	36.48	34.78	33.13	29.56	30.82	36.26
Qwen2.5-3B-Instruct	3	35.71	39.03	37.84	30.98	21.32	26.47	29.42	26.26	30.88
Qwen3-4B-Thinking-2507	4	43.52	55.22	50.08	44.92	37.10	39.55	38.57	35.84	43.10
Qwen3-4B-2507-Instruct	4	47.51	49.81	44.88	44.68	39.06	41.56	40.46	38.95	43.36
Qwen3.5-4B	4	46.15	46.93	44.57	41.72	35.11	37.00	35.37	31.55	39.80
DeepSeek-R1-Distill-Qwen-7B	7	38.06	39.66	41.38	35.33	26.23	30.90	31.74	28.89	34.02
Qwen2.5-7B-Instruct	7	47.30	45.71	44.00	42.95	33.38	35.77	39.05	33.29	40.18
DeepSeek-R1-Distill-Llama-8B	8	45.46	45.27	41.59	40.72	28.45	38.78	34.57	33.09	38.49
Llama-3.1-8B-Instruct	8	45.22	45.71	48.04	41.34	33.83	36.60	36.73	33.69	40.14
Qwen3-8B	8	43.15	53.15	48.65	48.45	38.36	42.11	40.33	39.37	44.20
Qwen3.5-9B	9	39.17	52.67	46.49	39.18	33.98	35.33	33.20	31.33	38.92
DeepSeek-R1-Distill-Qwen-14B	14	48.39	53.98	48.62	43.23	38.30	41.42	37.92	40.27	44.02
Qwen2.5-14B-Instruct	14	41.32	52.75	50.51	43.23	34.27	41.22	36.76	39.93	42.50
Qwen3-14B	14	44.89	48.24	50.52	47.50	38.49	42.68	39.63	37.73	43.71
Qwen3.5-27B	27	53.06	59.21	55.45	49.54	45.59	44.29	41.23	39.23	48.45
Qwen3.6-27B	27	50.91	55.30	49.88	47.69	38.50	43.25	42.28	38.63	45.80
Gemma-4-31B	30.7	52.52	57.78	50.98	49.30	42.41	46.25	43.08	43.89	48.28
Qwen2.5-32B-Instruct	32	41.33	51.75	53.26	42.88	33.36	45.24	39.45	40.01	43.41
QwQ-32B	32	47.17	49.77	54.41	45.88	37.42	41.62	38.17	36.23	43.83
Qwen3.5-Flash	35 / 3	51.93	56.08	52.10	52.38	41.68	45.06	42.34	37.98	47.44
Llama-3.1-70B	70	51.21	52.33	55.03	46.54	42.94	44.63	41.83	42.39	47.11
Qwen3-Next-80b-A3B-Instruct	80 / 3	50.01	55.26	52.59	50.65	42.03	45.00	41.15	38.45	46.89
Qwen3.5-122B-A10B	122 / 10	53.06	60.08	54.29	44.23	43.04	45.77	43.46	40.17	48.01
DeepSeek-V4-Flash	284 / 13	54.59	56.32	57.13	52.66	43.16	47.23	44.06	42.24	49.67
GLM-5.3-Flash	320 / 18	51.53	59.64	55.83	53.58	44.67	48.31	43.46	41.91	49.87
Qwen3.5-397B-A17B	397 / 17	52.55	59.21	54.82	52.49	44.18	46.82	42.53	41.19	49.22
DeepSeek-V3.2	671 / 37	50.65	54.04	52.88	46.70	38.26	43.01	39.81	36.92	45.28
DeepSeek-V4-Pro	1600 / 49	52.55	59.84	56.23	53.54	43.48	45.94	42.05	42.27	49.49

individual response prediction, motivating more explicit supervision of the predictive relationships between individual information and behavioral choices.

Table 6 compares 32 open weight models spanning 0.6B to 1600B total parameters across eight perception aware survey datasets. Average accuracy is computed by equally weighting the eight datasets.

Scaling Weakens Beyond 32B. Across the full model range, total parameter count is strongly associated with average accuracy, with Spearman's $\rho = 0.896$. However, for models with more than 32B total parameters, the correlation decreases to $\rho = 0.300$. Larger checkpoints therefore do not form a consistent performance ordering in individual response prediction. For example, DeepSeek-V4-Pro (1600B) achieves an average accuracy of 49.49%, slightly below DeepSeek-V4-Flash (284B) at 49.67%. Smaller models can also remain competitive on individual datasets, with Qwen3.5-27B and Gemma-4-31B achieving the highest accuracies on ISSP2015 and ISSP2021, respectively. Thus, although scaling is associated with substantial gains across the full model range, this relationship becomes much weaker among larger models.

From Model Scale to Decision Knowledge. All models in this comparison already receive demographic and perception information about each respondent. The comparison therefore holds the available individual information relatively rich while varying model capacity. What remains implicit is which observed factors are predictive of a particular response and how they should be connected to the final choice. Standard profile to option prediction provides the model with the individual information and the response target, but does not explicitly supervise these relationships.

The weakening scaling trend suggests that increasing model capacity alone does not reliably improve the use of this information for individual response prediction. SocialReasoner instead provides explicit supervision for these predictive relationships. Machine learning experts trained on survey

data identify evidence associated with correct predictions, semantic denoising removes unsupported evidence, and the retained information is converted into natural language reasoning traces for LLM training. In this way, SocialReasoner directly supervises the connection between observed individual information and behavioral choices rather than relying on model scale to acquire this knowledge implicitly.

These results therefore motivate shifting attention from model capacity alone to how decision knowledge is represented and supervised for human social reasoning.

B Additional Experimental Results and Ablations

B.1 Population-Level Survey Response Distribution Prediction

Summary. We examine whether improvements in individual response prediction also translate into more faithful population-level response distributions. Although SocialReasoner is trained only for individual prediction, its aggregated predictions substantially improve distribution fidelity over the base models. Across all three backbones, SocialReasoner also achieves lower cross-dataset average distances than SubPop on all four evaluation metrics, with improvements across most dataset–metric combinations. These results demonstrate that supervision grounded in individual decision evidence can improve both individual prediction and aggregate response patterns without an explicit distribution-level objective.

Constructing population response distributions. Beyond individual-level prediction accuracy, we evaluate whether aggregated model predictions reproduce the response distributions observed among survey respondents. We adopt hard aggregation, under which each respondent contributes one discrete predicted response. The human and model distributions are constructed from the same respondents for each question, allowing us to compare the response patterns induced by individual predictions with their empirical counterparts.

For survey question q, let N_q denote the number of respondents and K_q the number of response options. The empirical probability assigned to option k is

$$p_{qk} = \frac{1}{N_q} \sum_{n=1}^{N_q} \mathbf{1}\{y_{qn} = k\}, \tag{25}$$

and the corresponding model-predicted probability is

$$\hat{p}_{qk} = \frac{1}{N_q} \sum_{n=1}^{N_q} \mathbf{1}\{\hat{y}_{qn} = k\}. \tag{26}$$

Here, y_{qn} and $\hat{y}_{qn}$ denote the observed and predicted responses of respondent n, respectively. We denote the resulting distributions by $P_q = (p_{q1}, \ldots, p_{qK_q})$ and $\hat{P}_q = (\hat{p}_{q1}, \ldots, \hat{p}_{qK_q})$.

Evaluation metrics. We compare these distributions using total variation distance (TVD), Jensen–Shannon divergence (JSD), Hellinger distance (HD), and Wasserstein distance (WD). Lower values indicate closer agreement with the empirical human response distribution.

The total variation distance measures the discrepancy in probability mass assigned to the response options:

$$\mathrm{TVD}_q = \frac{1}{2} \sum_{k=1}^{K_q} |p_{qk} - \hat{p}_{qk}| \,. \tag{27}$$

Before computing JSD, we add $\epsilon = 10^{-12}$ to each probability and renormalize the distributions to avoid numerical issues caused by zero probabilities. Let P_q^ϵ and $\hat{P}_q^\epsilon$ denote the smoothed distributions, and let $M_q = (P_q^\epsilon + \hat{P}_q^\epsilon)/2$ denote their midpoint distribution. The Jensen–Shannon divergence is defined as

$$\mathrm{JSD}_q = \frac{D_{\mathrm{KL}}(P_q^\epsilon \| M_q) + D_{\mathrm{KL}}(\hat{P}_q^\epsilon \| M_q)}{2}. \tag{28}$$

The Hellinger distance measures the discrepancy between the square roots of the two probability distributions:

$$\mathrm{HD}_q = \sqrt{\frac{1}{2}\sum_{k=1}^{K_q}\left(\sqrt{p_{qk}} - \sqrt{\hat{p}_{qk}}\right)^2}. \tag{29}$$

For WD, the response options are placed in their encoded order at equally spaced positions on the normalized interval $[0, 1]$. Let $F_{qk} = \sum_{j=1}^{k} p_{qj}$ and $\hat{F}_{qk} = \sum_{j=1}^{k} \hat{p}_{qj}$ denote the corresponding cumulative distributions. The Wasserstein distance is computed as

$$\mathrm{WD}_q = \frac{1}{K_q - 1}\sum_{k=1}^{K_q-1}\left|F_{qk} - \hat{F}_{qk}\right|. \tag{30}$$

Each metric is computed separately for every survey question and then macro-averaged across all Q questions within the dataset:

$$\overline{D} = \frac{1}{Q}\sum_{q=1}^{Q} D_q. \tag{31}$$

Here, D_q is instantiated separately as TVD_q, JSD_q, HD_q, or WD_q. Thus, each dataset is evaluated using four distributional metrics, with every question receiving equal weight. The Average column in Table 7 further reports the unweighted mean of the eight dataset-level results for each metric.

Distribution fidelity across datasets and backbones. Table 7 compares the base models, SubPop, and SocialReasoner across eight survey datasets and three backbones. SocialReasoner improves over the corresponding base model on every reported dataset–metric combination. For example, with Qwen2.5-7B-Instruct, the average TVD decreases from 0.401 to 0.120, while the average JSD decreases from 0.143 to 0.018. The same pattern holds for Qwen3-4B-Instruct and Llama-3.1-8B-Instruct, showing that the distributional benefit extends across model families.

Compared with SubPop, which explicitly targets population response distributions, SocialReasoner achieves better results across most dataset–metric combinations. Its advantage is consistent at the cross-dataset average level: SocialReasoner obtains lower average TVD, WD, JSD, and HD under all three backbones. With Qwen2.5-7B-Instruct, the four average distances decrease from 0.128, 0.064, 0.024, and 0.155 under SubPop to 0.120, 0.047, 0.018, and 0.119 under SocialReasoner, respectively. These results show that learning to predict individual responses from decision-relevant evidence provides an effective route to population distribution recovery, even when the training objective does not explicitly align aggregate response frequencies.

Question-level response patterns. Figure 7 illustrates these differences using three randomly sampled ATP152 questions. The base model often concentrates its predictions on a narrow subset of options, producing response patterns that are more homogeneous than the empirical human distributions.

For the question about the impact of AI on truck-driving jobs, Qwen2.5-7B-Instruct assigns 86.2% of its predictions to "Fewer jobs," compared with 32.9% of human responses. SocialReasoner reduces this proportion to 30.2% and restores substantial probability mass to the other options. The TVD consequently decreases from 0.532 to 0.062. Similarly, for the question about AI making humans more productive, the base model assigns 78.8% of its predictions to "Somewhat likely," compared with 25.9% in the human data. SocialReasoner produces a more dispersed distribution, with 23.8% assigned to this option and a TVD of 0.094, compared with 0.528 for the base model. Across the examples, SocialReasoner better preserves both the diversity of selected responses and their relative frequencies.

From individual predictions to population patterns. Population response distributions summarize the choices of individual respondents. A model that repeatedly favors the same answer across heterogeneous profiles can therefore distort both individual predictions and aggregate response frequencies. SocialReasoner provides supervision on which respondent-specific observations support

Table 7: **Population-level response distribution fidelity across methods and datasets.** We compare hard-aggregated model predictions with empirical human response distributions using TVD, WD, JSD, and HD. Lower values indicate better distributional fidelity. Bold values denote the lowest reported value among methods for each model, dataset, and metric. When Ours ties for the lowest reported value, only Ours is bolded. Average is the unweighted arithmetic mean across the eight dataset values reported in this table, rounded to three decimal places.

Model	Method	Metric	ATP142	ATP152	EVS2017	ISSP2009	ISSP2015	ISSP2016	ISSP2019	ISSP2021	Average
Qwen2.5-7B-Instruct	Baseline	TVD	0.346	0.350	0.385	0.302	0.499	0.443	0.368	0.514	0.401
		WD	0.142	0.149	0.198	0.103	0.166	0.173	0.140	0.195	0.158
		JSD	0.107	0.108	0.139	0.093	0.197	0.163	0.128	0.210	0.143
		HD	0.324	0.335	0.372	0.313	0.456	0.416	0.365	0.472	0.382
	SubPOP	TVD	**0.128**	0.127	0.124	0.126	**0.129**	0.125	0.133	0.130	0.128
		WD	0.069	0.061	0.073	0.054	0.053	0.063	0.071	0.068	0.064
		JSD	**0.025**	0.021	0.026	0.021	**0.023**	0.022	0.026	0.025	0.024
		HD	0.167	0.146	0.147	0.149	**0.154**	0.148	0.167	0.162	0.155
	Ours	TVD	0.137	**0.103**	**0.112**	**0.092**	0.145	**0.116**	**0.128**	**0.128**	**0.120**
		WD	**0.062**	**0.043**	**0.043**	**0.030**	**0.053**	**0.052**	**0.044**	**0.052**	**0.047**
		JSD	0.026	**0.011**	**0.014**	**0.009**	0.030	**0.017**	**0.021**	**0.016**	**0.018**
		HD	**0.153**	**0.097**	**0.101**	**0.090**	0.158	**0.117**	**0.125**	**0.114**	**0.119**
Qwen3-4B-Instruct	Baseline	TVD	0.245	0.275	0.315	0.278	0.283	0.335	0.343	0.299	0.297
		WD	0.120	0.116	0.168	0.096	0.113	0.138	0.128	0.132	0.126
		JSD	0.060	0.068	0.098	0.069	0.076	0.107	0.105	0.082	0.083
		HD	0.223	0.243	0.298	0.248	0.260	0.310	0.306	0.272	0.270
	SubPOP	TVD	**0.111**	0.170	**0.111**	0.100	**0.102**	0.124	0.130	0.120	0.121
		WD	0.063	0.082	0.061	0.044	0.044	0.070	0.058	0.057	0.060
		JSD	**0.021**	0.033	0.017	0.015	0.016	0.022	**0.018**	0.015	0.020
		HD	**0.142**	0.184	0.119	0.123	0.124	0.151	0.132	0.122	0.137
	Ours	TVD	0.132	**0.113**	0.113	**0.090**	0.125	**0.117**	**0.127**	**0.116**	**0.117**
		WD	**0.060**	**0.047**	**0.045**	**0.030**	**0.040**	**0.051**	**0.046**	**0.048**	**0.046**
		JSD	0.024	**0.014**	**0.015**	**0.009**	**0.015**	**0.017**	0.020	**0.014**	**0.016**
		HD	0.145	**0.108**	**0.101**	**0.087**	**0.116**	**0.117**	**0.126**	**0.107**	**0.113**
Llama-3.1-8B-Instruct	Baseline	TVD	0.289	0.330	0.261	0.268	0.296	0.355	0.276	0.352	0.303
		WD	0.123	0.137	0.127	0.094	0.114	0.155	0.118	0.162	0.129
		JSD	0.079	0.090	0.069	0.063	0.084	0.116	0.078	0.118	0.087
		HD	0.273	0.292	0.253	0.254	0.296	0.333	0.287	0.345	0.292
	SubPOP	TVD	**0.136**	0.146	**0.116**	0.098	**0.105**	0.116	**0.098**	0.125	0.118
		WD	0.065	0.067	0.055	0.040	0.040	0.057	0.047	0.054	0.053
		JSD	**0.024**	0.025	**0.016**	0.014	**0.013**	0.018	**0.018**	0.023	0.019
		HD	0.161	0.160	0.118	0.121	**0.115**	0.134	0.142	0.154	0.138
	Ours	TVD	0.147	**0.103**	0.120	**0.097**	0.122	**0.113**	0.121	**0.111**	**0.117**
		WD	**0.063**	**0.041**	**0.044**	**0.031**	**0.039**	**0.050**	**0.043**	**0.046**	**0.045**
		JSD	0.027	**0.011**	0.017	**0.010**	0.016	**0.017**	0.019	**0.012**	**0.016**
		HD	**0.155**	**0.096**	**0.109**	**0.094**	0.117	**0.116**	**0.117**	**0.102**	**0.113**

a choice, encouraging the model to distinguish individuals through their decision-relevant information. The distributional results are consistent with this supervision helping the model recover variation that is suppressed by the base models.

Individual accuracy and distribution fidelity capture complementary aspects of performance on human social reasoning tasks. Matching aggregate option frequencies does not establish that the corresponding choices are assigned to the right individuals. Our results show that SocialReasoner improves both: the gains in individual response prediction reported in the main experiments are accompanied by more faithful population-level distributions. This joint improvement supports individual decision-evidence supervision as a useful foundation for modeling human social reasoning, with benefits that extend from respondent-level prediction to aggregate behavioral patterns.

B.2 Detailed Analysis of Machine Learning Models

Summary. We analyze seven machine learning models spanning additive, tree based, and neural architectures to understand the predictive structure available in survey data. Performance is highly consistent across model families but varies substantially across target questions, suggesting that predictability depends more on the information available for a question than on model architecture. Feature analyses further show that perception information dominates the most predictive respondent attributes, while removing perceptions reduces average prediction accuracy by 10.76 percentage points.

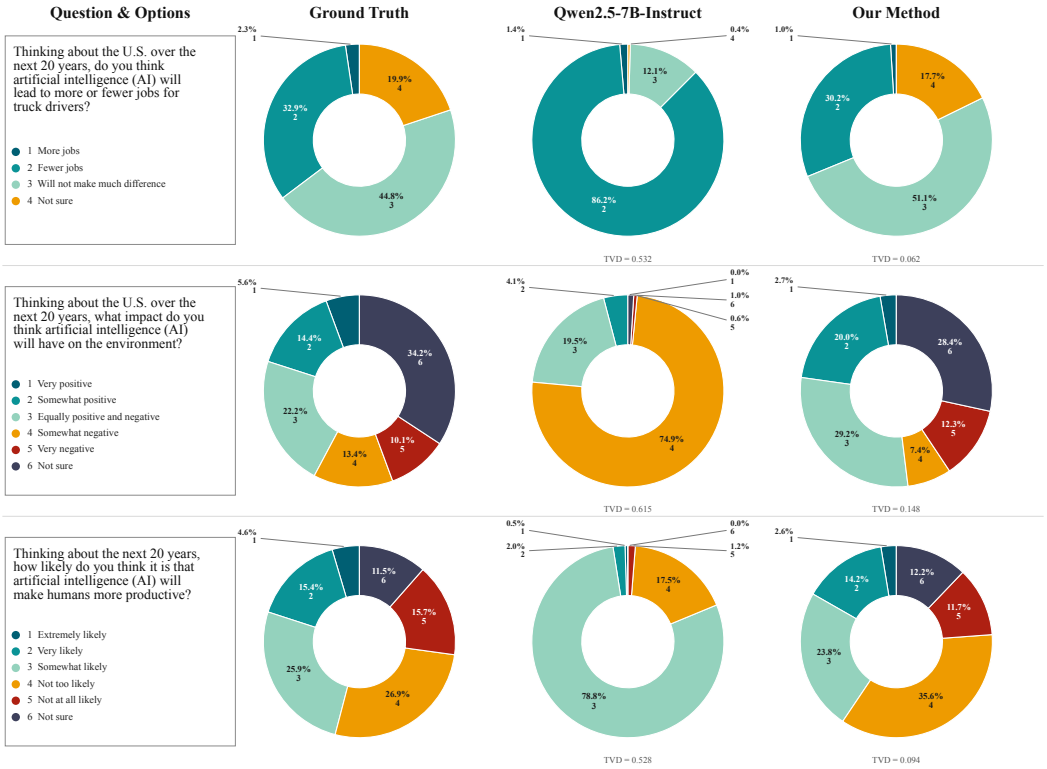

Figure 7: **Illustrative examples of option distributions on ATP152.** For three randomly sampled questions, we compare the empirical human response distribution (Ground Truth), predictions from Qwen2.5-7B-Instruct, and predictions from SocialReasoner. The base model concentrates its responses on a narrow subset of options, whereas SocialReasoner better preserves the diversity and relative frequencies of human responses. TVD is computed separately for each displayed question.

Experimental Settings. We evaluate seven representative machine learning models spanning three model families. The *linear models* include Logistic Regression and Explainable Boosting Machine (EBM) Nori et al. (2019); the *tree-based models* include Random Forest Breiman (2001), XGBoost Chen & Guestrin (2016), and LightGBM Ke et al. (2017); and the *neural models* include Multilayer Perceptron (MLP) and FT-Transformer Gorishniy et al. (2021). For EBM, we use a deliberately simplified additive configuration: all profile attributes are first one-hot encoded and feature interactions are disabled. Since each resulting input dimension is binary, the model reduces to an additive scoring function over individual encoded features, without cross-feature interaction terms. For each target survey question, we independently train a classifier using respondents from the corresponding training split and evaluate it on the held-out test respondents. Therefore, models are question-specific and do not share parameters across different survey questions. The detailed model architectures and training hyperparameters are summarized in Table 8.

Overall Performance across Datasets. Table 9 reports the individual-level prediction accuracy of the seven machine learning models across the eight survey datasets. We observe that performance varies substantially across datasets but remains remarkably consistent across model families within each dataset. For example, the average accuracy across models ranges from 54.47% on ISSP2021 to 64.00% on EVS2017, whereas the differences among models on the same dataset are generally small.

Tree-based models achieve a slight overall advantage, with XGBoost and LightGBM obtaining the highest average performance, but no single model consistently dominates across all datasets. Notably, the additive EBM remains competitive with substantially more complex tree-based and neural models, despite explicitly disabling feature interactions.

Table 8: **Hyperparameters of machine learning models.** The same configuration is used for each model across all datasets.

Linear Models		
Hyperparameter	**Logistic Regression**	**EBM**
Input encoding	One-hot	One-hot
Model form	Linear classifier	Additive main effects
Feature interactions	–	Disabled
Inverse L2 regularization strength	1.0	–
Optimizer	L-BFGS	Boosting
Learning rate	–	0.05
Maximum iterations / rounds	500	100
Minimum samples per leaf	–	4
Validation fraction	–	0.15

Tree-based Models			
Hyperparameter	**Random Forest**	**XGBoost**	**LightGBM**
Number of trees	500	300	300
Learning rate	–	0.05	0.05
Maximum tree depth	Unlimited	6	Unlimited
Minimum leaf / child size	5	1	20
Row sampling ratio	0.8	1.0	1.0
Feature sampling	$\sqrt{d}$	1.0	1.0
Number of leaves	–	–	31
L1 regularization	–	0	0
L2 regularization	–	1	0
Categorical smoothing	–	–	10

Neural Models		
Hyperparameter	**MLP**	**FT-Transformer**
Input representation	One-hot	Categorical embeddings
Hidden architecture	$(64, 32)$	2 Transformer layers
Token dimension	–	32
Attention heads	–	4
FFN expansion factor	–	2
Activation	ReLU	ReLU
Optimizer	AdamW	AdamW
Learning rate	1×10^{-3}	3×10^{-4}
Weight decay	1×10^{-4}	1×10^{-4}
Batch size	512	512
Maximum epochs	100	100
Validation fraction	0.15	0.15
Early-stopping patience	10	10
Dropout	–	0.1

These results suggest that much of the predictive signal contained in respondent profiles can be captured through relatively simple feature effects, while increasing model capacity or introducing more complex nonlinear interactions provides only modest additional gains.

Question-Level Prediction Performance. We further examine prediction performance for individual survey questions. As shown in Figure 8, prediction accuracy varies substantially across questions within the same dataset, while the seven machine learning models exhibit highly consistent patterns across questions. Questions that are difficult for one model tend to remain difficult for the others, whereas more predictable questions generally achieve high accuracy across all model fam-

Table 9: **Overall performance of machine learning models across datasets.** We report individual-level response prediction accuracy (%). The average denotes the mean performance of all seven models on each dataset.

Method	ATP142	ATP152	EVS2017	ISSP2009	ISSP2015	ISSP2016	ISSP2019	ISSP2021
Linear Models								
Logistic Regression	56.92	60.37	63.82	58.89	55.58	56.52	54.11	54.21
EBM	58.11	62.32	63.87	58.88	55.48	56.49	54.06	54.16
Tree-based Models								
Random Forest	58.09	62.42	63.41	58.84	55.43	56.29	54.23	54.03
XGBoost	57.19	62.24	64.61	59.65	56.13	57.59	54.97	55.09
LightGBM	56.93	61.94	64.01	59.81	56.02	57.86	55.21	55.20
Neural Models								
MLP	58.07	62.01	64.08	59.05	55.73	56.68	54.27	54.29
FT-Transformer	57.81	61.67	64.21	59.18	55.70	56.85	54.48	54.34
Average	**57.59**	**61.85**	**64.00**	**59.19**	**55.72**	**56.90**	**54.48**	**54.47**

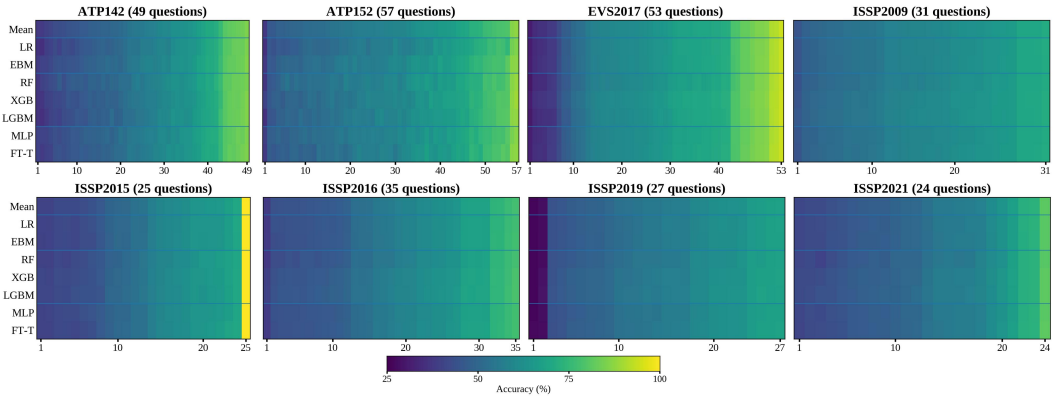

Figure 8: **Question-level prediction performance across datasets.** Each panel reports the individual-level prediction accuracy of seven machine learning models over all survey questions in one dataset. Questions are ordered from left to right by their mean accuracy across the seven models, from less predictable to more predictable. The top row in each panel shows the mean accuracy across models, while horizontal separators distinguish linear, tree-based, and neural model families.

ilies. This pattern is reflected by the pronounced vertical bands in the heatmaps, suggesting that variation in predictive performance is more closely associated with the target question than with the choice of model architecture. Together with the overall results in Table 9, these findings indicate that the predictability of individual human responses is shaped more strongly by the information available for a particular decision than by the expressive capacity of the predictive model. Increasing model complexity therefore does not substantially change which human responses can be reliably predicted from the available respondent information.

Feature Contributions to Human Response Prediction. To understand which types of respondent information contribute to human response prediction, we analyze feature contributions using permutation importance. For each model and survey question, we measure the decrease in prediction accuracy after randomly permuting one feature while keeping all other features unchanged. A larger accuracy drop indicates stronger predictive reliance on the corresponding feature. We then aggregate the importance scores across questions and across the seven machine learning models to obtain a unified measure of predictive importance for each respondent attribute.

Table 10: **Prediction performance without perceptual information.** Each cell reports accuracy (%), with the decrease from the corresponding full-input model shown in parentheses.

Model	ATP142	ATP152	EVS2017	ISSP2009	ISSP2015	ISSP2016	ISSP2019	ISSP2021
EBM	53.10 (↓5.01)	43.54 (↓18.78)	55.16 (↓8.71)	43.21 (↓15.67)	50.58 (↓4.90)	45.75 (↓10.74)	44.59 (↓9.47)	47.75 (↓6.41)
XGBoost	50.44 (↓6.75)	39.81 (↓22.43)	55.64 (↓8.97)	43.38 (↓16.27)	50.96 (↓5.17)	45.85 (↓11.74)	44.66 (↓10.31)	47.91 (↓7.18)
LightGBM	49.08 (↓7.85)	38.74 (↓23.20)	54.75 (↓9.26)	41.83 (↓17.98)	49.29 (↓6.73)	44.52 (↓13.34)	43.48 (↓11.73)	46.56 (↓8.64)
Logistic	52.91 (↓4.01)	42.80 (↓17.57)	55.24 (↓8.58)	43.23 (↓15.66)	51.09 (↓4.49)	45.77 (↓10.75)	44.66 (↓9.45)	47.79 (↓6.42)
Avg. Drop	**↓5.91**	**↓20.50**	**↓8.88**	**↓16.40**	**↓5.32**	**↓11.64**	**↓10.24**	**↓7.16**

As shown in Fig. 9, perceptual information consistently dominates the most predictive attributes across all eight datasets. Among the top 15 attributes ranked by aggregated importance, perceptual measurements account for 13 to 15 features in every dataset and contribute the majority of the total importance. This pattern remains stable across different survey programs and social domains, indicating that individual responses are more strongly associated with respondents' beliefs, attitudes, values, experiences, and related internal states than with conventional background characteristics.

Notably, our *Demographic & Context* category is deliberately broad. Many of the remaining predictive attributes in this category, such as political affiliation, religious affiliation, and volunteering behavior, already encode respondents' social orientation, value commitments, or patterns of social engagement. In contrast, purely static attributes such as age and gender are less prominent among the most predictive features. Therefore, even part of the signal captured by the non-perceptual group can be interpreted as a coarse proxy for latent preferences and behavioral tendencies. The dominance of perceptual information in Fig. 9 is thus obtained under a relatively conservative grouping scheme.

This observation also helps explain why machine learning models are useful for constructing reasoning traces in our framework. The predictive models are optimized to identify respondent information that best discriminates among possible decisions. Because perceptual attributes provide stronger predictive signals, the resulting expert-derived signal naturally assigns greater importance to beliefs, attitudes, values, and experiences. When these predictive patterns are converted into reasoning traces, the generated reasoning therefore tends to emphasize cognitively relevant evidence rather than relying primarily on demographic associations. Importantly, this behavior does not require manually specifying which perceptual cues should be used for each question; it emerges from their predictive value in the observed human responses.

We further verify this effect by removing all perceptual measurements and training the same machine learning models using only demographic and contextual information. Table 10 reports the resulting accuracy together with the decrease relative to the corresponding full-input model. Removing perceptual information causes an average accuracy reduction of **10.76 percentage points** across the four evaluated models and eight datasets. The reduction is particularly large on ATP152 and ISSP2009, reaching 20.49 and 16.40 percentage points on average across models, respectively. Although demographic and contextual attributes retain non-trivial predictive information, these results show that they recover only a limited portion of the signal available from the complete respondent representation. Together with the permutation analysis, the results provide consistent evidence that perceptual information constitutes the primary predictive source for modeling human social reasoning.

B.3 Robustness Across Model Scales

Summary. We further evaluate SocialReasoner with Llama-3.1-70B-Instruct (Grattafiori et al., 2024) to examine whether its effectiveness extends to a substantially larger backbone. On ATP152, SocialReasoner improves response prediction accuracy from 52.33% to 63.51%, corresponding to an absolute gain of 11.18 percentage points. This result shows that the benefit of decision knowledge transfer is not limited to the smaller backbones evaluated in the main experiments.

The main experiments evaluate SocialReasoner with models ranging from 4B to 8B parameters. To examine robustness at a larger model scale, we additionally apply the same framework to Llama-3.1-70B-Instruct (Grattafiori et al., 2024) and evaluate it on ATP152.

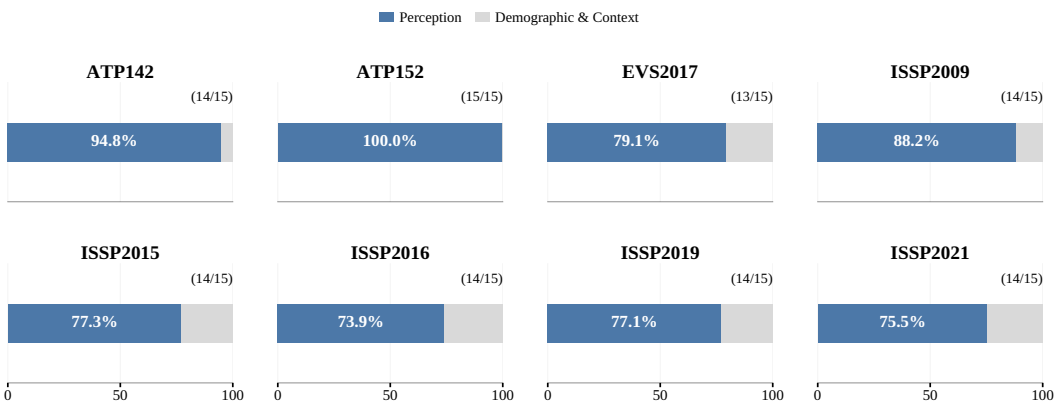

Figure 9: **Perceptual information dominates the most predictive respondent attributes across datasets.** We aggregate permutation importance across seven machine learning models, rank all respondent attributes by their aggregated importance, and examine the top 15 features. The blue segment denotes the proportion of aggregated importance contributed by perceptual information, while the gray segment represents demographic and contextual information. The notation $(x/15)$ indicates the number of perceptual features among the top 15 most predictive attributes.

Table 11: Performance of SocialReasoner with a larger backbone on ATP152.

Backbone	Method	Accuracy (%) ↑
Llama-3.1-70B-Instruct	Base	52.33
Llama-3.1-70B-Instruct	**SocialReasoner**	**63.51**

As shown in Table 11, the 70B base model achieves 52.33% accuracy, while SocialReasoner reaches 63.51%, yielding an 11.18 percentage point improvement. The magnitude of this gain remains substantial despite the much larger backbone capacity.

Together with the results on Qwen3-4B, Qwen2.5-7B, and Llama-3.1-8B in the main experiments, this result indicates that SocialReasoner remains effective across a broad range of model scales. In particular, increasing the capacity of the base model does not eliminate the benefit of explicitly transferring decision knowledge from structured behavioral data. SocialReasoner can therefore complement stronger backbone models rather than serving only as a mechanism for improving smaller LLMs.

B.4 Efficiency Analysis

Summary. SocialReasoner substantially improves individual response prediction without relying on long reasoning traces. On ISSP2009, it achieves 59.41% accuracy while using 65.3% fewer output tokens and 63.8% lower response latency than standard CoT prompting. The additional LoRA parameters require only 0.1562 GB of memory. These results show that the gains of SocialReasoner come with substantially lower reasoning and inference costs than conventional CoT, supporting concise, evidence grounded supervision for efficient human social reasoning.

Table 12 compares the predictive performance and inference efficiency of the baseline, CoT, and our lightweight LoRA-based method on ISSP2009. The baseline achieves an accuracy of 41.34% with the lowest response latency of 0.086 seconds and only 2.0 output tokens on average, as it directly predicts the answer without an explicit reasoning process. In contrast, CoT substantially increases the inference cost, generating 386.6 output tokens and requiring 4.433 seconds per response on average, while its accuracy decreases to 37.38%. This observation suggests that the effectiveness of CoT does not lie in producing longer reasoning traces. Rather, what matters is whether the reasoning process can accurately identify, integrate, and prioritize the information that is truly relevant to an individual's decision.

Table 12: Efficiency and accuracy comparison on ISSP2009.

Method	Accuracy (%) ↑	TPS ↑	TPR (s) ↓	Add. Param. Mem. (GB) ↓	Output Tokens ↓
Baseline	41.34	23.382	**0.086**	**0.0000**	**2.0**
CoT	37.38	**87.211**	4.433	**0.0000**	386.6
Ours	**59.41**	83.451	1.606	0.1562	134.0

Our lightweight method achieves the highest accuracy of 59.41%, outperforming the baseline and CoT by 18.07 and 22.03 percentage points, respectively. More importantly, this improvement does not rely on excessively long reasoning. Compared with CoT, our method reduces the average number of output tokens from 386.6 to 134.0, corresponding to a reduction of 65.3%, while decreasing the average response latency from 4.433 seconds to 1.606 seconds, a reduction of 63.8%. Meanwhile, its decoding throughput remains comparable to CoT, with 83.451 versus 87.211 tokens per second, while requiring only 0.1562 GB of additional LoRA parameter memory. These results indicate that effective human social reasoning is not determined by the length of the reasoning chain. Instead, the key lies in learning an accurate, concise, and decision-relevant reasoning process. Our method therefore provides a lightweight approach that substantially improves prediction accuracy while maintaining considerably lower reasoning and inference costs.

B.5 Effect of Evidence-Grounded Supervision

Summary. We isolate the contribution of evidence supervision while keeping the training records, instance multiplicities, and optimization settings fixed. Answer only supervision already achieves strong performance, showing that expert-guided record construction provides a useful training signal. However, random evidence offers little additional benefit, while LLM selected evidence mainly improves cross survey transfer. SocialReasoner achieves the best performance on both ATP152 and GSS2022, reaching 63.22% and 50.78%, respectively. These results show that both the expert-induced training distribution and the content of the reasoning supervision contribute to performance, with predictive evidence learned from behavioral data providing additional value beyond semantic relevance alone.

SocialReasoner affects both the composition of the training data and the reasoning content used for supervision. Only instances supported by correct expert predictions are retained, and the same respondent–question instance may appear multiple times when valid traces are constructed from evidence provided by several experts. We therefore construct controlled variants that preserve the same training records and isolate the contribution of evidence supervision.

All variants are constructed from exactly the same final SocialReasoner training records. We preserve the task inputs, response labels, multiplicity of each respondent–question instance, backbone model, fine tuning configuration, and number of optimization updates. The variants differ only in the supervision presented to the student.

We consider four settings. *Answer only* removes the reasoning trace and supervises only the final response while preserving the records and multiplicities produced by the expert-guided construction pipeline. *Random evidence* retains the same records but replaces the SocialReasoner evidence with randomly sampled non-target facts from the same respondent profile. For each record, we match the number of evidence statements in the corresponding SocialReasoner trace and keep the supervision length as close as possible without modifying the sampled facts. *LLM selected evidence* also retains the same records, but replaces ML based evidence selection with facts selected by an LLM according to their semantic relevance to the target question. The LLM only selects existing facts from the non-target respondent profile and does not rewrite them, generate additional explanations, or determine the final answer. Finally, *SocialReasoner* uses evidence identified from predictive relationships learned by the ML experts and retained after semantic denoising.

All models are trained only on ATP152 using Qwen2.5-7B-Instruct as the backbone. We evaluate them on the ATP152 test set and directly on GSS2022 without further adaptation. ATP152 measures performance on the source survey, while GSS2022 evaluates transfer to an unseen survey.

Answer only reaches 60.82% on ATP152 despite containing no explicit reasoning trace. Since its training records and multiplicities are inherited from the expert-guided construction pipeline, this

Table 13: Effect of different supervision signals on source survey prediction and cross survey transfer. All variants use identical ATP152 training records, instance multiplicities, and optimization updates. GSS2022 is used only for evaluation.

Supervision	**ATP152** ↑	**GSS2022** ↑
Answer only	60.82	49.10
Random evidence	61.43	49.21
LLM selected evidence	61.54	50.41
SocialReasoner	**63.22**	**50.78**

result shows that the ML experts already provide useful supervision by determining which instances are retained and how frequently they appear. Their predictive knowledge is therefore partly encoded in the resulting training distribution even before evidence is expressed in natural language.

The evidence controls further show that the content of the reasoning supervision matters. Random evidence improves ATP152 accuracy only modestly, from 60.82% to 61.43%, and produces almost no change on GSS2022 (49.10% to 49.21%). LLM selected evidence achieves a similar ATP152 result of 61.54%, but improves GSS2022 accuracy to 50.41%. This suggests that semantic relevance alone contributes relatively little to source survey prediction, but provides more useful information when transferring to an unseen survey.

SocialReasoner achieves the strongest performance in both settings, reaching 63.22% on ATP152 and 50.78% on GSS2022. Compared with Answer only, this corresponds to gains of 2.40 and 1.68 percentage points, respectively. More importantly, SocialReasoner improves ATP152 accuracy by 1.68 points over LLM selected evidence, showing that semantic relevance alone does not recover the predictive information identified from behavioral data. On GSS2022, LLM selected evidence already recovers much of the transfer benefit, while SocialReasoner provides a further 0.37 point improvement.

Taken together, these results separate the effects of training record construction and evidence content. Expert-guided record construction provides an implicit predictive signal through the training distribution. Adding arbitrary respondent information contributes little, while semantically relevant evidence is particularly useful for cross survey transfer. SocialReasoner further grounds this evidence in predictive relationships learned from behavioral data, yielding the strongest performance on both the source survey and the unseen survey.

B.6 Individual Expert Supervision and Repeated Exposure

Summary. The three individual experts yield similar student performance, indicating that SocialReasoner is not driven by a single dominant expert. Repeating the supervision from one expert three times provides a modest improvement, showing that increased training exposure contributes to performance. However, even after matching the training scale, all repeated single-expert variants remain below the full multi-expert SocialReasoner model. These results suggest that combining supervision from different predictive models provides an additional benefit beyond repeated exposure alone.

We further examine how individual machine learning experts contribute to SocialReasoner. The default framework combines valid reasoning traces constructed from evidence provided by EBM, LightGBM, and MLP, representing additive, tree based, and neural predictive models, respectively. In the main experiments, progressively combining traces constructed from expert evidence improves downstream performance, but also increases the amount of training supervision. We therefore construct controlled variants to separate the contribution of the expert source from the effect of increased training exposure.

We first train separate student models using traces constructed from the evidence of only one expert. For each variant, the complete SocialReasoner trace construction pipeline remains unchanged, including out of fold prediction, local evidence extraction, and semantic denoising. The student receives valid traces constructed from the evidence of EBM, LightGBM, or MLP alone, while the full SocialReasoner setting combines all valid traces derived from the evidence of the three experts.

Table 14: Effect of supervision derived from individual experts and repeated exposure on student accuracy on ATP152. The repeated variants match the training scale of SocialReasoner by repeating the same traces constructed from one expert's evidence three times without introducing new evidence.

Supervision Source	# Training Records	Accuracy (%) ↑
Individual Expert Supervision		
EBM	121.7k	62.31
LightGBM	122.3k	62.44
MLP	120.7k	62.49
Repeated Single Expert Supervision		
EBM $\times 3$	365.1k	62.62
LightGBM $\times 3$	366.9k	62.79
MLP $\times 3$	362.1k	62.78
SocialReasoner	**364.7k**	**63.22**

The trace construction pipeline yields similar numbers of valid training records for the three experts: approximately 121.7k for EBM, 122.3k for LightGBM, and 120.7k for MLP. Combining the three sources produces approximately 364.7k records. To control for this difference in training scale, we additionally repeat each single-expert corpus three times without introducing any new reasoning trace or evidence. This produces 365.1k, 366.9k, and 362.1k records for EBM, LightGBM, and MLP, respectively, closely matching the 364.7k records used by SocialReasoner. These variants therefore increase exposure to the same supervision while keeping the set of unique traces constructed from that expert's evidence fixed.

All variants use Qwen2.5-7B-Instruct with the same fine tuning configuration and one epoch training protocol, and are evaluated on the same ATP152 test set.

The student models trained separately on supervision derived from the three experts achieve closely matched performance, ranging from 62.31% to 62.49%. The small variation across the EBM-, LightGBM-, and MLP-derived supervision settings indicates that the effectiveness of SocialReasoner is not attributable to one particularly strong supervision source. Instead, each expert provides useful evidence for supervision despite relying on a different predictive architecture.

Repeating a single-expert corpus three times improves student accuracy in every case. With EBM-derived supervision, accuracy increases from 62.31% to 62.62%; with LightGBM-derived supervision, from 62.44% to 62.79%; and with MLP-derived supervision, from 62.49% to 62.78%. Thus, increased exposure to the same supervision accounts for part of the gain observed when the training corpus becomes larger.

Repeated exposure alone, however, does not reproduce the performance of multi-expert supervision. The strongest repeated single-expert variant reaches 62.79%, while SocialReasoner achieves 63.22% with a closely matched number of training records. Since the repeated variants contain comparable training volume but no additional unique reasoning traces, the remaining improvement cannot be explained solely by the number of optimization examples.

Together, these results indicate that both training exposure and expert combination contribute to SocialReasoner. Repeated supervision from a single expert provides a modest benefit, while combining traces constructed from evidence provided by different predictive models yields an additional improvement at comparable training scale. This supports the use of multiple experts as complementary sources of supervision rather than simply as a mechanism for increasing the amount of training data.

B.7 Effect of Reinforcement Learning and Data Selection

Summary. Most of the performance gain of SocialReasoner is obtained during supervised fine tuning, which improves Qwen2.5-7B-Instruct from 45.71% to 63.22% on ATP152. Additional reinforcement learning provides a smaller improvement, and its effectiveness depends strongly on the training data used for RL. Standard GRPO on the full training set slightly decreases accuracy to 62.89%, whereas GRPO restricted to instances correctly predicted by the machine learning experts

Table 15: Effect of reinforcement learning and RL data selection on ATP152. All GRPO variants start from the same SocialReasoner SFT checkpoint.

Training Setting	**Accuracy (%) ↑**
Base model	45.71
SocialReasoner SFT	63.22
SocialReasoner SFT + GRPO (Full)	62.89
SocialReasoner SFT + GRPO (ML unsolved)	60.87
SocialReasoner SFT + GRPO (ML solved)	**63.83**

further improves performance to 63.83%. In contrast, training only on expert unsolved instances reduces accuracy to 60.87%.

Many LLM post training pipelines apply reinforcement learning after supervised fine tuning. We therefore examine whether RL can provide additional gains after SocialReasoner supervision. We use Qwen2.5-7B-Instruct on ATP152. The base model achieves 45.71% accuracy, while SocialReasoner supervised fine tuning improves performance to 63.22%. Starting from this same SFT checkpoint, we then apply standard GRPO under different choices of RL training data.

In addition to using the full training set, we partition the behavioral instances according to whether the machine learning experts correctly predict the observed response. *ML solved* contains instances correctly predicted by the experts, while *ML unsolved* contains the remaining instances. We apply the same GRPO procedure using the full training set, the ML solved subset, or the ML unsolved subset.

As shown in Table 15, the dominant improvement comes from supervised fine tuning. SocialReasoner SFT raises accuracy from 45.71% to 63.22%, an absolute gain of 17.51 percentage points. In comparison, the best RL setting provides a further gain of 0.61 points, reaching 63.83%. These results indicate that RL serves as an additional refinement stage rather than the primary source of improvement.

The benefit of RL, however, depends strongly on which behavioral instances are used for optimization. Applying GRPO to the full training set slightly reduces accuracy from 63.22% to 62.89%. Restricting RL to ML unsolved instances results in a larger decrease to 60.87%. By contrast, GRPO using only ML solved instances improves accuracy to 63.83%, outperforming both SocialReasoner SFT and full data GRPO.

This contrast suggests that not all behavioral examples provide equally useful supervision for reinforcement learning. Expert correctness offers an empirical criterion for selecting instances whose observed respondent information supports the recorded response under predictive models trained on the structured data. In our experiments, these instances provide a more effective RL signal than either the complete dataset or the expert unsolved subset.

We do not interpret ML unsolved instances as inherently noisy or unpredictable. Rather, they may contain weaker predictive relationships under the available information or patterns that are not captured by the current machine learning experts. The results therefore support a narrower conclusion: after SocialReasoner supervised fine tuning, reinforcement learning can provide an additional improvement, but this improvement depends on selecting behavioral instances that provide suitable optimization signals.

Overall, supervised fine tuning accounts for the majority of SocialReasoner's performance gain, while reinforcement learning provides a smaller but complementary improvement when the RL data are appropriately selected.

C Method Details

C.1 Feature Attribution for Decision Evidence Construction

SocialReasoner uses local attribution to make the decision relationships captured by heterogeneous experts explicit. Because different expert paradigms use different input representations and attri-

Table 16: **Local attribution rules for decision evidence construction.** Each expert assigns a local support score to observed respondent information. Scores are used only for ranking within the same expert and prediction instance and are not compared numerically across expert families.

Expert	Input representation	Explained output	Local support score
EBM	One-hot encoded fields with additive main effects	Raw class score	Additive term associated with the observed category
LightGBM	Original categorical fields	Raw class score	TreeSHAP value of the corresponding field
Random Forest	One-hot encoded fields	Class probability	Sum of TreeSHAP values over the encoded block of the field
MLP	One-hot encoded fields	Class logit	Expected marginal logit change induced by the complete field block
FT-Transformer	Categorical field identifiers	Class logit	Expected marginal logit change induced by the categorical field

bution mechanisms, their attribution values are not assumed to share a common numerical scale. Instead, each expert independently ranks the observed respondent information according to its local support for a given behavioral choice. These rankings are then mapped back to the original respondent information to construct candidate decision evidence.

For an accepted out-of-fold prediction, the same fitted expert is used for both behavioral prediction and local attribution. Let

$$\mathbf{v} = (v_1, \ldots, v_H)$$

denote the respondent fields available to the current instance, and let

$$a = \hat{y}_{i,j}^{(m)} = y_{i,j}$$

denote the correctly predicted behavioral choice. We define

$$\alpha_h^{(m)}(\mathbf{v}; a)$$

as the local support score assigned by expert m to the observed value of field h for choice a. This corresponds to the local attribution $c_{i,j,h}^{(m)}$ defined in Section 3.3, with the instance indices omitted here for clarity.

The precise definition of $\alpha_h^{(m)}$ depends on the expert architecture. Importantly, these scores are used only to rank information within the same expert and prediction instance. Attribution values from different expert families are never directly compared or numerically aggregated. Table 16 summarizes the corresponding representations and attribution rules. EBM, LightGBM, and MLP constitute the default heterogeneous experts used by SocialReasoner, while Random Forest and FT-Transformer are additionally considered in the expert diversity analyses.

EBM. The Explainable Boosting Machine is fitted on one-hot encoded respondent fields with feature interactions disabled, yielding an additive scoring function over the encoded indicators. Let $e_r(\mathbf{v}) \in \{0, 1\}$ denote encoded indicator r, and let $r(h, v_h)$ identify the indicator corresponding to the observed value v_h of field h. For behavioral choice a, the EBM score is

$$s_a^{\mathrm{EBM}}(\mathbf{v}) = b_a + \sum_{r=1}^{R} f_{r,a}(e_r(\mathbf{v}))\,, \tag{32}$$

where b_a is the class-specific intercept, R is the number of encoded indicators, and $f_{r,a}$ is the learned additive function for indicator r.

For decision evidence construction, the support associated with the observed value of field h is

$$\alpha_h^{\mathrm{EBM}}(\mathbf{v}; a) = f_{r(h,v_h),a}(1). \tag{33}$$

This score represents the local support associated with the observed category, rather than a complete decomposition of the EBM prediction. This distinction is intentional: the active indicator corresponds directly to an observed respondent fact that can be expressed in natural language, whereas inactive indicators correspond to alternative categories that were not observed. The complete EBM score in Eq. 32 is always used for prediction; Eq. 33 is used only to rank observed information for evidence construction.

For binary decisions, the signed contribution is oriented toward the accepted choice so that positive values always indicate support for that choice.

LightGBM. LightGBM operates directly on the original categorical respondent fields. For each accepted prediction, TreeSHAP decomposes the raw score of the accepted choice into field-level contributions and a bias term. The bias term is excluded from decision evidence construction. Let $\phi_{h,a}^{\mathrm{LGBM}}(\mathbf{v})$ denote the TreeSHAP contribution of field h for choice a. We define

$$\alpha_h^{\mathrm{LGBM}}(\mathbf{v}; a) = \phi_{h,a}^{\mathrm{LGBM}}(\mathbf{v}). \tag{34}$$

Because LightGBM operates on the original fields, no aggregation over encoded indicators is required. For binary decisions, the signed raw margin is oriented toward the accepted choice using the same convention as for EBM.

Random Forest. Random Forest uses one-hot encoded respondent fields. TreeSHAP attribution is first computed at the encoded-column level and then aggregated back to the corresponding original field. Let G_h denote the set of encoded columns derived from field h, and let $\phi_{r,a}^{\mathrm{RF}}(\mathbf{v})$ denote the TreeSHAP value of encoded column r for choice a. The local support score is

$$\alpha_h^{\mathrm{RF}}(\mathbf{v}; a) = \sum_{r \in G_h} \phi_{r,a}^{\mathrm{RF}}(\mathbf{v}). \tag{35}$$

The aggregation covers the complete encoded block associated with field h. The resulting score is then mapped back to the respondent's observed value of that field. This differs from EBM, where the active observed category itself defines the unit used for evidence ranking.

MLP and FT-Transformer. For neural experts, we estimate field-level support using marginal changes in the logit of the accepted choice under background replacement. The attribution unit is the original respondent field rather than an individual encoded dimension.

For each respondent, we sample background profiles from the training fold of the corresponding target question. Starting from a background profile, respondent fields are introduced according to sampled field orders. For MLP, introducing a field replaces its complete one-hot block. For FT-Transformer, it replaces the corresponding categorical field identifier.

Let $\mathbf{v}^{(b)}$ denote background profile b, and let $\mathbf{v}_S^{(b)}$ denote the hybrid profile in which fields in S take their values from the respondent being explained while the remaining fields retain their values from background b. For sampled order p, let $S_{b,p,h}$ denote the set of fields introduced before field h. The marginal contribution of field h is

$$\Delta_{b,p,h}^{(m)}(a) = s_a^{(m)}\left(\mathbf{v}_{S_{b,p,h}\cup\{h\}}^{(b)}\right) - s_a^{(m)}\left(\mathbf{v}_{S_{b,p,h}}^{(b)}\right), \tag{36}$$

where $m \in \{\mathrm{MLP}, \mathrm{FT\text{-}Transformer}\}$. Averaging over B sampled backgrounds and P sampled field orders gives

$$\alpha_h^{(m)}(\mathbf{v}; a) = \frac{1}{BP}\sum_{b=1}^{B}\sum_{p=1}^{P}\Delta_{b,p,h}^{(m)}(a). \tag{37}$$

This quantity estimates the expected marginal support contributed by field h across different contexts, avoiding reliance on a single masking or deletion operation. For MLP, we use eight background profiles and eight sampled field orders per background. For FT-Transformer, we use four

backgrounds and four sampled field orders. Along each complete replacement path, the sum of successive marginal changes recovers the difference between the respondent and background logits, which we use as a numerical consistency check.

From Local Attribution to Decision Evidence. After obtaining expert-specific local support scores, we map the highest-ranked fields back to the respondent information explicitly present in the original input. Missing values, duplicated facts, and fields that cannot be reliably grounded in the rendered respondent profile are excluded.

The mapping preserves the identity and observed value of each original measurement and excludes the target item itself. Thus, every candidate fact passed to semantic denoising is traceable to information explicitly available to the LLM rather than to an interpretation generated by the expert.

Following the primary rule in Section 3.3, we retain at most K_e observed facts with positive support for the accepted choice and rank them by $\alpha_h^{(m)}(\mathbf{v}; a)$. Ranking is performed independently for every expert and prediction instance, so attribution scales from different expert families do not interact.

Alternative Attribution Criteria. We additionally consider two alternative ways of converting local attribution into candidate decision evidence.

The first uses a *contrastive support score*, measuring how strongly a field favors the accepted choice over its strongest competing choice. For multiclass experts,

$$\alpha_{h,\mathrm{ctr}}^{(m)}(\mathbf{v}; a) = \alpha_{h,a}^{(m)}(\mathbf{v}) - \max_{a' \neq a} \alpha_h^{(m)}(\mathbf{v}; a) \tag{38}$$

Only fields with positive contrastive support are retained. For binary experts represented by a single signed margin, the oriented scalar directly provides the corresponding contrastive score.

The second criterion retains both positive and negative support for the accepted choice. Positive values represent supporting evidence, whereas negative values are preserved as counterevidence rather than reversing their direction. This allows the resulting reasoning supervision to reflect locally conflicting signals rather than presenting every selected fact as support for the final choice.

Together, these attribution procedures provide a common interface between heterogeneous decision experts and natural language social reasoning: each expert exposes locally relevant relationships in the structured behavioral data, while the subsequent semantic denoising stage determines which of these relationships provide a defensible basis for the target decision. The semantic denoising and reasoning construction procedures are described in Appendix C.2.

C.2 The Role of LLM Intervention in Social Reasoning Construction

The heterogeneous experts recover relationships between individual information and behavioral choices from structured behavioral data. These relationships must then be expressed in a form that LLMs can learn as human social reasoning. We therefore examine how different forms of LLM intervention affect this process. The six variants range from directly preserving expert evidence to semantic filtering, evidence synthesis, reasoning reconstruction, and independent decision making. Together, these comparisons clarify what knowledge should be supplied by the experts and what role the LLM should play in expressing the underlying social reasoning. All reported accuracies in this section are evaluated on ATP142 using Qwen2.5-7B-Instruct as the backbone.

No LLM Denoising. We first consider a setting in which the LLM does not participate in semantic refinement. For each retained expert prediction, the selected respondent information is converted into atomic natural language statements and paired with the accepted behavioral choice. This preserves the relationships identified by the experts without additional semantic filtering.

Statistical relevance, however, does not necessarily provide a meaningful basis for human reasoning. A factor may contribute to an expert prediction through correlations in the behavioral data while having little semantic relevance to the target decision. Once verbalized as reasoning, such a factor may be presented as explanatory evidence despite the absence of a clear connection to the corre-

sponding human choice. The following examples illustrate this distinction and show how semantic denoising changes the evidence exposed to the LLM.

From Expert Selected Evidence to Denoised Reasoning

Example 1: K–12 STEM Education

Question: Compared with other wealthy nations, how does the U.S. compare on science, technology, engineering and math education in grades K to 12?

Expert Selected Evidence:

- The respondent rates the U.S. health care system as below average relative to other wealthy nations.
- The respondent reports census division as "Pacific".
- The respondent rates the effect of access to grocery stores on a person's weight as some.
- The respondent reports education as "some college".

After LLM Denoising:

✓ U.S. health care system is rated below average relative to other wealthy nations.

× Pacific census division.

× Perceived effect of grocery store access on body weight.

× Some college education.

The health care evaluation provides a related assessment of U.S. institutional performance within the same comparative frame. The remaining evidence has a less direct semantic connection to the respondent's evaluation of STEM education.

Example 2: U.S. Entertainment

Question: Compared with other wealthy nations, how does the U.S. compare on entertainment, including movies, music, and television?

Expert Selected Evidence:

- The respondent rates the U.S. health care system as the best relative to other wealthy nations.
- The respondent rates the U.S. military as the best relative to other wealthy nations.
- The respondent reports income as 90,000 to less than 100,000.
- The respondent reports being extremely comfortable discussing mental health with a close friend.

After LLM Denoising:

✓ U.S. health care system is rated the best relative to other wealthy nations.

✓ U.S. military is rated the best relative to other wealthy nations.

× Household income.

× Comfort discussing mental health with a close friend.

The denoiser retains related evaluations of U.S. performance while removing personal attributes with a less evident connection to the entertainment judgment.

Example 3: Conflicting but Relevant Evidence

Question: Compared with other wealthy nations, how does the U.S. compare on entertainment, including movies, music, and television?

Observed Choice: The U.S. is above average.

Expert Selected Evidence:

- The respondent rates U.S. K–12 STEM education as below average relative to other wealthy nations.
- The respondent rates the U.S. military as above average relative to other wealthy nations.
- The respondent reports being very comfortable discussing mental health with a religious or spiritual leader.
- The respondent reports being not too concerned about COVID-19 hospitalization.

After LLM Denoising:

- ✓ U.S. K–12 STEM education is rated below average relative to other wealthy nations.
- ✓ U.S. military is rated above average relative to other wealthy nations.
- × Comfort discussing mental health with a religious or spiritual leader.
- × Concern about COVID-19 hospitalization.

The STEM evaluation is retained even though its direction differs from the final choice. This example illustrates that semantic relevance does not require agreement with the observed behavior. The observed choice is shown only for interpretation and is not provided to the denoiser.

Without denoising, accuracy reaches 57.28%, compared with 47.30% for the base model. The large improvement shows that relationships recovered by the experts already provide useful supervision. Semantic denoising addresses a different question: which statistically informative signals provide a meaningful basis for the corresponding human decision?

LLM Denoising. We use GLM-5.3-Flash (Zeng et al., 2026) as a constrained semantic denoiser. It receives the target question, answer options, and expert selected evidence, but neither the observed behavioral choice nor the expert prediction. Its role is to identify irrelevant or redundant information and remove evidence whose apparent relevance depends on unsupported group stereotypes. It does not introduce new evidence or determine the final decision.

Each candidate fact is classified as retained or removed. If no evidence remains, the corresponding instance is excluded from reasoning supervision. This design restricts the LLM to judging semantic relevance while keeping the source of the decision relationships grounded in structured behavioral data.

LLM Denoising improves accuracy from 57.28% to 58.27%, achieving the best performance among the six variants. As illustrated above, the denoiser can preserve relevant but conflicting evidence rather than selecting only information that agrees with the final choice. These results support using the LLM to refine decision relationships recovered by the experts rather than allowing it to reconstruct the decision process independently.

LLM Summary. We next examine whether further synthesizing the refined evidence produces stronger social reasoning supervision. The LLM receives both the retained evidence and the accepted expert decision. It combines related observations, organizes their relationships, and expresses an overall judgment connecting the evidence to the fixed behavioral choice. Thus, the decision remains grounded in the expert prediction, while the LLM has greater freedom in expressing how the evidence is integrated.

The following example illustrates this process. Related observations about religious belief are combined into a broader interpretation, while nominal religious affiliation is considered together with the respondent's explicitly stated beliefs and priorities.

Example of LLM Summary

Generated Reasoning:

The respondent explicitly identifies as a convinced atheist, reports no belief in God, and describes their conception of God as no spirit, God, or life force. Consistent with this, they rate religion as not important in their life, do not believe in an afterlife, heaven, or hell, and would not choose religious faith as a quality for children to learn at home. Although nominally affiliated with a religious denomination, their stated beliefs and the low value they place on religion indicate that God plays no meaningful role for them, so they would most likely say God is not at all important in their life. Therefore, the answer is 1.

Analysis. The summary combines several related observations about disbelief and the low importance assigned to religion. It also distinguishes religious affiliation from substantive belief, explaining why the former does not outweigh the respondent's explicit statements. These observations are then integrated into an overall judgment supporting the supplied choice.

LLM Summary achieves 57.87% accuracy, slightly below the 58.27% obtained by LLM Denoising. Although summarization makes the connection between the observed information and final decision more explicit, the additional synthesis does not improve the downstream learning signal. This result

suggests that semantically refined evidence already provides an effective basis for learning social reasoning, without requiring a more elaborate explanation.

LLM Rewriting. This variant retains the accepted expert decision but removes the expert selected evidence. The LLM is asked to construct reasoning that supports the supplied choice. The final behavioral decision therefore remains grounded in a correct expert prediction, but the reasoning must be reconstructed from the LLM's pretrained knowledge rather than from the relationships recovered from behavioral data.

This setting isolates the value of the decision from the value of the evidence underlying it. A generated explanation may be compatible with the supplied answer without recovering the individual information that supported the original expert prediction. Correctness of the final choice therefore does not imply that the reasoning captures the corresponding human decision process.

LLM Rewriting achieves 55.80% accuracy, compared with 57.87% for LLM Summary and 58.27% for LLM Denoising. The performance gap highlights the importance of grounding reasoning in explicit decision relationships rather than providing the final decision alone. We also observe less stable loss reduction during training, consistent with the greater variation introduced when reasoning is generated without expert evidence.

LLM Decision-Making. We further examine the complementary setting in which the expert selected evidence is retained but the accepted expert decision is withheld. The LLM interprets the evidence, independently selects an option, and generates the corresponding reasoning. We apply rejection sampling and retain only trajectories whose final choices match the observed behavior.

This variant reverses LLM Rewriting. Rewriting supplies the decision without the expert evidence, whereas LLM Decision-Making supplies the evidence but delegates both its interpretation and the final decision to the LLM. Rejection sampling guarantees correctness of the retained decisions, but does not guarantee that the inferred reasoning reflects the relationships recovered from structured behavioral data.

LLM Decision-Making achieves 54.23% accuracy, below the variants that preserve both the accepted expert decision and its supporting evidence. The result suggests that allowing the LLM to reconstruct the decision logic from selected information does not strengthen the resulting social reasoning supervision.

Hybrid LLM CoT. Finally, we combine the LLM's native reasoning with reasoning grounded in expert evidence. For each training instance, the LLM first selects an answer and generates reasoning without expert assistance. If its answer matches the observed behavior, the resulting trajectory is retained. Otherwise, expert selected evidence is provided and rejection sampling is used to obtain a correct trajectory. The resulting training corpus therefore combines native LLM reasoning with reasoning grounded in structured behavioral data.

This strategy achieves 51.52% accuracy, the lowest among the six variants, and also exhibits less stable loss reduction during training. One possible explanation is that the source of reasoning becomes dependent on instance difficulty. Instances already solved by the LLM rely on its pretrained social knowledge, whereas harder instances rely on relationships recovered by the experts. This heterogeneous source of reasoning does not provide a stronger learning signal than a consistent construction grounded in expert evidence.

Discussion. These comparisons clarify the complementary roles of heterogeneous experts and the LLM in SocialReasoner. The experts recover relationships between individual information and observed behavior, while the LLM is most effective when restricted to assessing the semantic relevance of these relationships. Providing the decision without its supporting evidence, delegating the decision to the LLM, or mixing native LLM reasoning with expert grounded reasoning all reduce performance. Additional synthesis after semantic denoising also provides no further gain.

Together, these findings support the central design of SocialReasoner: the knowledge underlying human social reasoning should remain grounded in relationships recovered from structured behavioral data, while the LLM serves primarily to express and semantically refine this knowledge in natural language.

D Detailed Discussion of Related Works

LLMs for Social Science Applications With the rapid development of large language models, a growing body of work has explored their applications in economic, financial, and broader social science settings, particularly for prediction and decision support. LLMs have been applied to diverse financial tasks, including question answering, information extraction, text analysis, market forecasting, and decision making (Xie et al., 2024). This progress has also motivated the development of domain specific financial language models such as BloombergGPT and Xuanyuan 2.0, which improve financial capabilities through domain adaptation, supervised fine tuning, or their combination (Wu et al., 2023; Zhang et al., 2023). More recent studies further enhance financial reasoning through specialized reasoning data and reinforcement learning (Liu et al., 2025b; Zhu et al., 2025).

Prior work has also shown that textual signals extracted by LLMs can provide useful predictive information for financial markets (Lopez-Lira & Tang, 2023), demonstrating their potential to capture knowledge relevant to real world decisions. At the same time, many social science prediction problems rely on abundant structured data in which the intermediate knowledge connecting observed characteristics to final decisions is rarely expressed in natural language corpora. Structured behavioral data therefore provide an important complementary source of decision knowledge beyond text and direct input output supervision. Our work explores this direction by extracting predictive evidence from structured human response data and transferring it to LLMs through explicit reasoning supervision.

LLM-Based Agent Simulation of Social Systems LLM agents have been increasingly used to simulate complex social systems across law, economics, healthcare, finance, recommendation, and other domains. AgentCourt (Chen et al., 2025) models legal procedures through agents with distinct judicial roles, while EconAgent (Li et al., 2024b) studies macroeconomic dynamics through interacting agents. WarAgent (Hua et al., 2023) simulates historical interstate conflicts, and AgentHospital (Li et al., 2024a) models medical workflows and clinical decision making.

LLM agents have also been applied to economic and financial systems. FinCon (Yu et al., 2024) studies collective investment decisions through networks of investment agents, while MarketSim (Piao et al.) simulates stock markets through interactions among heterogeneous generative agents. Related work further shows that pricing agents can exhibit supracompetitive behavior in oligopolistic settings (Fish et al., 2024). Beyond these domains, Agent4Rec (Zhang et al., 2024a) simulates user behavior in recommender systems, while ElectionSim (Zhang et al., 2024b) explores large scale electoral simulation.

Several general purpose platforms have further expanded the scope of LLM based social simulation, including AgentSociety (Piao et al., 2026; 2025), YulanSim (Wang et al., 2025c), and SOTOPIA-S4 (Zhou et al., 2025). These systems support interactive and large scale simulations across diverse social settings. Many of them instantiate individuals primarily through role descriptions or personal profiles. Such information provides useful background context, but relatively static profile attributes may not capture the perceptions, beliefs, and evaluations most relevant to a particular decision. Our work complements these systems by focusing on the fidelity of the underlying individual model and by incorporating perception information together with evidence grounded decision modeling.

LLM-Based Simulation of Individual Behavior Simulating individual human behavior with LLMs has become an important direction in computational social science and provides a behavioral basis for broader social simulation. Existing studies evaluate whether model-generated decisions reproduce patterns observed in human experiments and surveys. At the aggregate level, LLMs have shown considerable promise. For example, recent work finds that they can predict treatment effects in social science experiments with performance comparable to pooled human forecasters (Ashokkumar et al., 2026). Such results demonstrate that LLMs can capture meaningful regularities in collective human behavior.

Aggregate fidelity, however, does not necessarily imply faithful simulation at the individual level. Even when population-level patterns are reproduced, model responses may remain overly homogeneous and fail to capture variation across individuals (Mei et al., 2024; Hu et al., 2025). This limita-

tion is particularly important for social simulation, where realistic population behavior depends not only on average tendencies but also on heterogeneous decisions among individuals.

Prior work further shows that LLMs can systematically deviate from observed human decision patterns. In particular, models often produce choices that are more consistent with normative rationality than those made by human participants (Liu et al., 2024). Explicit Chain-of-Thought prompting and reasoning-oriented models can amplify this tendency, shifting decisions further toward expected value maximization rather than the behavior observed in human data (Liu et al., 2024; Tak et al., 2026). Thus, stronger general reasoning does not necessarily translate into better modeling of human social reasoning.

Recent benchmarks based on large scale human experiments and sociological surveys confirm that substantial gaps remain in predicting individual responses across tasks and populations (Hu et al., 2025; Wang et al., 2025b). Together, these findings suggest that the central challenge is not simply to elicit more reasoning from LLMs, but to provide the individual specific information and decision knowledge needed to reproduce heterogeneous human choices. Our work addresses this challenge by incorporating observed perception information and explicitly supervising the intermediate mapping from respondent specific evidence to behavioral responses.

Aligning LLMs with Human Responses Recent studies have explored several complementary approaches to improving LLMs' performance on human social reasoning tasks. SOCRATES and CENTAUR adapt language models directly on human response data, bringing model predictions closer to observed human choices (Kolluri et al., 2025; Binz et al., 2025). SUBPOP and DSA instead align model outputs with the response distributions of different human populations, improving distributional fidelity and individual response prediction (Suh et al., 2025; Cao et al., 2025; Huang et al., 2026). SAFO further develops supervised adaptation for human behavioral data (Lin et al., 2026), while HARPO uses reinforcement learning to better capture heterogeneity across human responses (Ong et al., 2026). Other methods, including CLAIMSIM and COO, improve performance on human social reasoning tasks through prompting strategies that encourage models to organize and evaluate available evidence without additional training (Yu et al., 2025; Do et al., 2025).

These approaches improve performance on human social reasoning tasks through response supervision, distribution alignment, prompting, or optimization over behavioral objectives. SocialReasoner focuses on a complementary source of supervision. Rather than relying only on final responses or reasoning produced by the language model itself, we use predictive machine learning experts to extract respondent specific evidence from structured behavioral data. This evidence is then converted into reasoning supervision that explicitly connects observed individual information to the final choice. In this way, SocialReasoner transfers decision knowledge that remains implicit in structured human response data into a unified language model.

Reasoning Supervision from Structured Data Intermediate reasoning has also been explored as an additional source of supervision for language model adaptation. DISTILLING STEP-BY-STEP uses rationales generated by a large language model as additional supervision, showing that intermediate explanations can provide richer learning signals than answer labels alone (Hsieh et al., 2023). More recently, TLRD extends this idea to tabular data by transforming label-only datasets into structured rationale supervision that integrates instance features, dataset statistics, and retrieved examples (Liang et al., 2026).

SocialReasoner shares the broader goal of exposing information that remains implicit in structured data, but differs in how the reasoning supervision is obtained. Rather than asking a language model to construct the decision process directly, SocialReasoner first trains predictive machine learning experts on the observed data and extracts respondent specific evidence from their correct predictions. The language model is then used only to remove evidence that lacks a defensible semantic relationship with the target question. As a result, the reasoning supervision is primarily grounded in predictive relationships learned from structured behavioral data, while limiting reliance on reasoning generated from the language model's own prior knowledge.

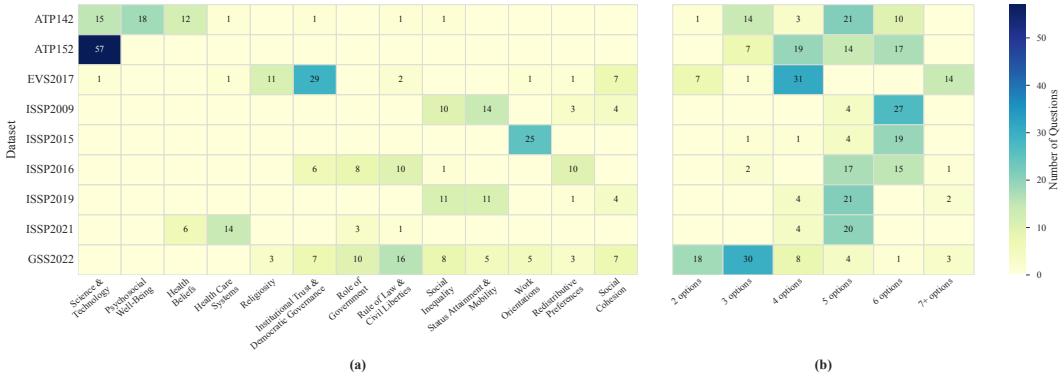

Figure 10: **Topical coverage and response structure across datasets.** (a) Distribution of questions across social-science areas. (b) Distribution of questions by response-option cardinality. Cell values indicate the number of questions in each category.

Table 17: **Overview of the survey datasets used in our experiments.** Areas denotes the number of thematic areas covered by each dataset, and Instances denotes the number of valid respondent–question pairs. GSS2022 is used for out-of-distribution evaluation.

Dataset	Coverage	Areas	Respondents	Questions	Instances
ATP142	U.S.	7	9,851	49	330,148
ATP152	U.S.	1	5,227	57	244,976
EVS2017	36 countries	8	50,916	53	2,552,997
ISSP2009	41 countries	4	48,822	31	1,503,093
ISSP2015	36 countries	1	45,096	25	682,115
ISSP2016	35 countries	5	43,439	35	1,501,834
ISSP2019	29 countries	4	40,245	27	1,044,243
ISSP2021	30 countries	4	37,899	24	841,611
GSS2022	U.S.	9	500	64	24,649
Overall	–	**13**	**281,995**	**365**	**8,725,666**

E Datasets Details

Although existing benchmarks such as SocialBench, SocioBench, and OptionQA have advanced the evaluation of human social reasoning in LLMs Chen et al. (2024); Wang et al. (2025b); Santurkar et al. (2023), they largely overlook the role of human perceptions in shaping individual behavior. To address this limitation, we construct a new benchmark incorporating perception information from four established large-scale survey programs, namely the American Trends Panel (ATP), the European Values Study (EVS), the International Social Survey Programme (ISSP), and the General Social Survey (GSS). In total, we derive nine datasets spanning multiple countries and regions and covering a broad range of social-science domains. An overview of the survey programs and datasets is provided in Table 17. Figure 10 further summarizes the composition of the benchmark by showing the distribution of response-option cardinalities across social-science areas within each survey program.

E.1 American Trends Panel

American Trends Panel Wave 142 (ATP142). Wave 142 of the Pew Research Center's American Trends Panel was fielded from February 7 to 11, 2024 Pew Research Center (2024a). We retain 49 target questions spanning seven social science domains, including Science & Technology, Psychosocial Well-Being, Health Beliefs, Social Inequality, Rule of Law & Civil Liberties, Health Care Systems, and Institutional Trust & Democratic Governance. The resulting dataset contains 330,148 valid respondent–question instances from 9,851 U.S. respondents. We split the respondents into 7,881 for training and 1,970 for testing, with no respondent overlap between the two splits.

Example question. *How much trust do you have in the information that comes from ChatGPT about the 2024 U.S. presidential election?*

Options. (1) A great deal; (2) Quite a bit; (3) Some; (4) Not too much; (5) None at all; (6) Not sure.

American Trends Panel Wave 152 (ATP152). Wave 152 of the Pew Research Center's American Trends Panel was fielded from August 12 to 18, 2024 Pew Research Center (2024b). We retain 57 target questions in the Science & Technology domain, covering AI awareness and attitudes, perceived control over AI, anticipated societal and labor-market impacts, comparisons between human and AI capabilities, trust and perceived risks, and views on AI governance and regulation. The resulting dataset contains 244,976 valid respondent–question instances from 5,227 U.S. respondents. We split the respondents into 4,182 for training and 1,045 for testing, with no respondent overlap between the two splits.

Example question. *Thinking about the U.S. over the next 20 years, what impact do you think artificial intelligence (AI) will have on medical care?*

Options. (1) Very positive; (2) Somewhat positive; (3) Equally positive and negative; (4) Somewhat negative; (5) Very negative; (6) Not sure.

E.2 European Values Study

European Values Study 2017 (EVS2017). EVS2017 is derived from the fifth wave of the European Values Study using the EVS 2017 Integrated Dataset, with country-specific fieldwork conducted between 2017 and 2021 EVS (2022). We retain 53 target questions spanning eight social science domains, including Religiosity, Institutional Trust & Democratic Governance, Social Cohesion, Rule of Law & Civil Liberties, Redistributive Preferences, Health Care Systems, Science & Technology, and Work Orientations. The questions primarily examine religious beliefs, interpersonal and institutional trust, political attitudes, democratic governance, social cohesion, and preferences concerning public institutions and social arrangements. The resulting dataset contains 2,552,997 valid respondent–question instances from 50,916 respondents across 36 countries. We split the respondents into 40,732 for training and 10,184 for testing, with no respondent overlap between the two splits.

Example question. *How important is religion in your life?*

Options. (1) Very important; (2) Quite important; (3) Not important; (4) Not at all important.

E.3 International Social Survey Programme

ISSP Social Inequality IV 2009 (ISSP2009). ISSP2009 is derived from the Social Inequality IV module of the International Social Survey Programme, with fieldwork conducted across participating national and regional samples between 2008 and 2012 ISSP Research Group (2017a). We retain 31 target questions spanning four social science domains, including Status Attainment & Mobility, Social Inequality, Redistributive Preferences, and Social Cohesion. The questions examine perceived determinants of socioeconomic advancement, inequality and redistribution, fairness in access to social resources, and perceived conflict across social groups. The resulting dataset contains 1,503,093 valid respondent–question instances from 48,822 respondents across 41 national or regional samples. We split the respondents into 39,060 for training and 9,762 for testing, with no respondent overlap between the two splits.

Example question. *How important do you think coming from a wealthy family is for getting ahead in life?*

Options. (1) Essential; (2) Very important; (3) Fairly important; (4) Not very important; (5) Not important at all; (8) Can't choose / Don't know.

ISSP Work Orientations IV 2015 (ISSP2015). ISSP2015 is derived from the Work Orientations IV module of the International Social Survey Programme, with fieldwork conducted across participating national and regional samples between 2015 and 2017 ISSP Research Group (2017b). We retain 25 target questions in the Work Orientations domain, covering work values and job characteristics, preferences over working hours and earnings, job commitment and mobility, employment

insecurity and expectations, attitudes toward trade unions and older workers, and willingness to make trade-offs between working hours and earnings. The resulting dataset contains 682,115 valid respondent–question instances from 45,096 respondents across 36 national or regional samples. We split the respondents into 36,064 for training and 9,032 for testing, with no respondent overlap between the two splits.

Example question. *How important is high income to you personally?*

Options. (1) Very important; (2) Important; (3) Neither important nor unimportant; (4) Not important; (5) Not important at all; (8) Can't choose.

ISSP Role of Government V 2016 (ISSP2016). ISSP2016 is derived from the Role of Government V module of the International Social Survey Programme (ZA6900, Version 2.0.0), with fieldwork conducted across participating national and regional samples between 2015 and 2018 ISSP Research Group (2018). We retain 35 target questions spanning five social science domains, including Rule of Law & Civil Liberties, Role of Government, Redistributive Preferences, Social Inequality, and Institutional Trust & Democratic Governance. The questions cover government responsibilities and spending priorities, taxation and redistribution, political efficacy and institutional influence, corruption and accountability, and attitudes toward civil liberties, surveillance, and national security. The resulting dataset contains 1,501,834 valid respondent–question instances from 43,439 respondents across 35 national or regional samples. We split the respondents into 34,751 for training and 8,688 for testing, with no respondent overlap between the two splits.

Example question. *Should it be the government's responsibility to provide decent housing for those who cannot afford it?*

Options. (1) Definitely should be; (2) Probably should be; (3) Probably should not be; (4) Definitely should not be; (8) Can't choose / Don't know.

ISSP Social Inequality V 2019 (ISSP2019). ISSP2019 is derived from the Social Inequality V module of the International Social Survey Programme, with fieldwork conducted across participating national samples between 2018 and 2022 ISSP Research Group (2022). We retain 27 target questions spanning four social-science domains, including Status Attainment & Mobility, Social Inequality, Redistributive Preferences, and Social Cohesion. The questions cover perceived determinants of socioeconomic advancement, income and wealth inequality, redistribution and taxation, fairness in access to social resources, perceived conflict between social groups, social mobility expectations, and generalized trust. The resulting dataset contains 1,044,243 valid respondent–question instances from 40,245 respondents across 29 national samples. We split the respondents into 32,196 for training and 8,049 for testing, with no respondent overlap between the two splits.

Example question. *To what extent do you agree or disagree with the statement: Differences in income in [COUNTRY] are too large?*

Options. (1) Strongly agree; (2) Agree; (3) Neither agree nor disagree; (4) Disagree; (5) Strongly disagree.

ISSP Health and Health Care II 2021 (ISSP2021). ISSP2021 is derived from the Health and Health Care II module of the International Social Survey Programme, with fieldwork conducted across participating national samples between 2021 and 2024 ISSP Research Group (2024). We retain 24 target questions spanning four social science domains, including Health Care Systems, Health Beliefs, Role of Government, and Rule of Law & Civil Liberties. The questions cover confidence in and evaluations of healthcare systems, inequalities in access to care, trust in physicians, beliefs about the causes and prevention of health problems, public provision of health care, and government authority during public-health emergencies. The resulting dataset contains 841,611 valid respondent–question instances from 37,899 respondents across 30 national samples. We split the respondents into 30,318 for training and 7,581 for testing, with no respondent overlap between the two splits.

Example question. *How much confidence do you have in the health care system in your country?*

Options. (1) Complete confidence; (2) A great deal of confidence; (3) Some confidence; (4) Very little confidence; (5) No confidence at all.

Table 18: **Hyperparameters for LoRA fine-tuning.**

Hyperparameter	Value
Optimizer	AdamW
Learning rate	2×10^{-4}
Weight decay	0
Warmup ratio	0.03
Training epochs	1
Effective batch size	16
Maximum sequence length	4,096
LoRA rank r	32
LoRA α	32
LoRA dropout	0.05

E.4 General Social Survey

General Social Survey 2022 (GSS2022). GSS2022 is derived from the 2022 General Social Survey conducted by NORC at the University of Chicago, with fieldwork conducted from May 4 to December 20, 2022 Davern et al. (2024). We retain 64 target questions spanning nine social science domains, including Religiosity, Institutional Trust & Democratic Governance, Role of Government, Rule of Law & Civil Liberties, Social Inequality, Status Attainment & Mobility, Work Orientations, Redistributive Preferences, and Social Cohesion. The resulting dataset contains 24,649 valid respondent–question instances from 500 U.S. respondents. All 500 respondents are used for testing, with no respondent overlap with any of the training datasets.

Example question. *Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?*

Options. (1) Can trust; (2) Can't be too careful.

F Details On Experimental Settings

F.1 Decoding and Sampling Settings

Following SocioBench Wang et al. (2025b), we use a temperature of 0.6, top-p of 0.9, and top-k of 50 for response generation. Unless otherwise specified, the maximum generation length is 512 tokens, which is used for all standard models and methods, including SocialReasoner.

The only exception is the evaluation of long-reasoning models in Appendix A.6. We found that a 512-token limit frequently truncates their reasoning before a final response is produced. For these models, we therefore increase the maximum generation length to 10,000 tokens while keeping the remaining decoding settings unchanged. This larger budget is used only to evaluate reasoning models under their intended long-form generation behavior and is not used by SocialReasoner.

F.2 Training Hyperparameters

Following Centaur Binz et al. (2025), we adapt all backbone models using Low-Rank Adaptation (LoRA) Hu et al. (2021) rather than full-parameter fine-tuning. By restricting parameter updates to a small set of low-rank parameters while keeping the pretrained backbone frozen, LoRA provides a parameter-efficient way to adapt general-purpose language models to human social reasoning. This design also limits changes to the pretrained model parameters, helping preserve the capabilities acquired during pretraining and mitigate catastrophic forgetting. Unless otherwise specified, we use the same training configuration across all fine-tuned backbone models. The main hyperparameters are summarized in Table 18.

F.3 Prompt Details

Standard Evaluation Prompt.

Instruction
Answer the survey question as this respondent. Return only the option number.

Survey Context
Survey fieldwork period and temporal context

Personal Information
Respondent demographic and socioeconomic information

Attitudinal and Psychological Measurement Profile
These measurements are derived from this respondent's answers to other non-target survey items. They capture the respondent's beliefs, attitudes, values, perceived risks, trust, confidence, preferences, experiences, and expectations.

Non-target survey items and corresponding responses

Question
Target survey question

Options
1. Option 1
2. Option 2
...
K. Option K

Requirements
Respond with the number only.

Chain of Thought.

Requirements
Please reason step by step about which option this human respondent would choose.
Your "reason" should carefully explain the rationale for the choice.
Your "option" must contain only the number of the chosen option.
Do not output any text outside the JSON object.

Few-Shot Neighbor.

Similar Respondent Answer Reference
A training respondent with the most similar personal profile answered the same question as follows. Use this answer only as a weak reference signal; still predict the current respondent's own answer.

Similarity score

Similar respondent's answer

Evidence Filtering.

You are a strict, answer-blind cue pruner for a survey rationale. Retain all cues by default and omit a cue only if it meets one of the following conditions:

Irrelevant cue. The cue has no defensible direct relationship to the target question.

Stereotype-dependent cue. Relating the cue to the question would require an unsupported inference based on demographic or group identity.

Duplicate cue. The cue expresses the same respondent information as another cue without adding distinct information.

Target question and response options

Ordered candidate cues

Do not infer the answer. Do not rewrite, summarize, connect, or reorder the cues. Retain all cues unless one of the above conditions clearly applies.

F.4 HARDWARE CONFIGURATION

All experiments were conducted on a server equipped with eight NVIDIA A100-SXM4-80GB GPUs, each with 80 GB of GPU memory, to ensure a consistent hardware environment across model training and evaluation.

G CASE STUDY

We present case studies across eight datasets to illustrate the connections between respondents' perceptions, SocialReasoner-generated reasoning, and the predicted responses. Highlighted spans identify the evidence discussed in the accompanying author analysis.

Case Study: Psychosocial Well-Being ATP142

Prompt: Instruction: Answer the survey question as this respondent. Return only the option number.
Survey Context: This survey was conducted from February 7 to February 11, 2024. Interpret words such as today, current conditions, and references to the 2024 U.S. presidential election relative to the survey date.
Demographic Profile: Age group: 65+; Birthplace: U.S. - 50 states, District of Columbia; Census division: Middle Atlantic; Country: United States; Education level: secondary or vocational; Hispanic identity: no; Income: $100,000 or more; Household income position: upper; Ideology: moderate; Internet use: several times a day; Marital status: married; Metro status: non-metropolitan; Party leaning: dem/Lean Dem; Race/ethnicity: White non-Hispanic; Region: Northeast; Religion: catholic; Religious attendance: once a week; Sex: female; Residential area: rural; Volunteer activity: no; Voter registration: you are ABSOLUTELY CERTAIN that you are registered to vote at your current address; Years in U.S.: born in US (not including unincorporated territories).
Perceptions: These measurements are derived from this respondent's answers to other non-target survey items. They capture the respondent's beliefs, attitudes, values, perceived risks, trust, confidence, preferences, experiences, and expectations. Assessment of U.S. universities relative to other wealthy nations: **The U.S. is the best**; Assessment of the U.S. standard of living relative to other wealthy nations: **The U.S. is above average**; Assessment of the U.S. military relative to other wealthy nations: **The U.S. is above average**; Assessment of the U.S. health care system relative to other wealthy nations: **The U.S. is above average**; Assessment of the U.S. political system relative to other wealthy nations: **The U.S. is above average**; Assessment of U.S. scientific achievements relative to other wealthy nations: **The U.S. is above average**; Preferred level of government regulation of major technology companies: **More than they are now**; View on whether major technology companies favor liberals or conservatives: **Conservatives over liberals**; Perceived likelihood that social media companies censor political viewpoints: **Not very likely**; General view on whether generative AI systems should credit their sources: **Need to credit the sources they reviewed to come up with their responses because they rely on work done by others**; View on whether generative AI should credit sources when using the writing style of a deceased author: **AI programs should have to credit their sources if they do this**; View on whether generative AI should credit sources when using images in the style of a current artist: **Not sure**; View on whether generative AI should credit sources when using a summary of U.S. population information: **Not sure**; Perceived threat of COVID-19 to U.S. public health: **A major threat**; Perceived threat of COVID-19 to the U.S. economy: **Not a threat**; Importance of understanding and treating Long COVID: **Very important**; Importance of keeping up with friends and family to personal well-being: **Very important**; Importance of regular exercise to personal well-being: **Very important**; Importance of spending time outdoors to personal well-being: **Very important**; Importance of having a fulfilling job to personal well-being: **Not too important**; Importance of spiritual or religious practices to personal well-being: **Extremely important**; Comfort discussing mental health with a close friend: **Somewhat comfortable**; Comfort discussing mental health with a spouse or partner: **Somewhat comfortable**; Comfort discussing mental health with an immediate family member: **Somewhat comfortable**; Comfort discussing mental health with a mental health therapist: **Very comfortable**; Comfort discussing mental health with a neighbor: **Not too comfortable**; Perceived effect of genetic factors on a person's weight: **Some**; Perceived effect of access to grocery stores on a person's weight: **Some**; Belief about willpower and long-term weight loss: **Willpower alone is usually not enough to lose weight and keep it off**; Perceived threat of COVID-19 to personal health: **A major threat**; Perceived threat of COVID-19 to personal finances: **Not a threat**.
Question: How important is taking time to relax for your own overall well-being?
Options: 1. Extremely important; 2. Very important; 3. Somewhat important; 4. Not too important; 5. Not at all important.
Requirements: Respond with the number only.

SocialReasoner-generated CoT

<think>The respondent considers spending time outdoors very important to personal well-being.
The respondent considers regular exercise very important to personal well-being.
The respondent considers keeping up with friends and family very important to personal well-being.
The respondent reports being very comfortable discussing mental health with a mental health therapist.
Taken together, these ordered profile details are most consistent with "Very important". Therefore, the answer is 2.</think>

Analysis

The importance assigned to exercise, outdoor activity, and social connection suggests an orientation toward maintaining well-being through everyday practices. Interpreting relaxation as another such practice connects these preferences to the target judgment, providing a coherent basis for the "Very important" prediction.

Prediction: 2 — Very important
Observed response: 2 — Very important

Figure 11: SocialReasoner case study on psychosocial well-being (ATP142).

Case Study: Science & Technology ATP152

Prompt: Instruction: Answer the survey question as this respondent. Return only the option number.
Survey Context: This survey was conducted from August 12 to August 18, 2024. Interpret words such as today and time periods such as the next 20 years relative to the survey date.
Demographic Profile: Age group: 50-64; Birthplace: U.S. - 50 states, District of Columbia; Census division: West South Central; Country: United States; Education level: associate degree; Hispanic identity: no; Income: $60,000 to less than $70,000; Household income position: middle; Ideology: very conservative; Internet use: several times a day; Marital status: married; Metro status: metropolitan; Party leaning: rep/Lean Rep; Race/ethnicity: White non-Hispanic; Region: South; Religion: protestant; Religious attendance: more than once a week; Sex: male; Residential area: rural; Volunteer activity: no; Voter registration: you are ABSOLUTELY CERTAIN that you are registered to vote at your current address; Years in U.S.: born in US (not including unincorporated territories).
Perceptions: These measurements are derived from this respondent's answers to other non-target survey items. They capture the respondent's beliefs, attitudes, values, perceived risks, trust, confidence, preferences, experiences, and expectations. Overall feeling about AI in daily life: **More concerned than excited**; Perceived control over how AI is used in personal life: **Not too much**; Preferred level of control over how AI is used in personal life: **I'd like more control over how AI is used in my life**; Expected overall impact of AI on the U.S. over the next 20 years: **Very negative**; Expected effect of AI on the number of U.S. jobs: **Fewer jobs**; Expected personal impact of increased AI use: **Harm you than benefit you** (prompt rendering: More harm than benefit); Assessment of the current level of government regulation of AI: **Not go far enough regulating its use** (prompt rendering: Does not go far enough); Expected impact of AI on medical care: **Very negative**; Expected impact of AI on K-12 education: **Very negative**; Expected impact of AI on elections: **Very negative**; Expected impact of AI on the economy: **Very negative**; Expected impact of AI on the criminal justice system: **Somewhat negative**; Expected effect of AI on jobs for lawyers: **Fewer jobs**; Expected effect of AI on jobs for software engineers: **Will not make much difference**; Expected effect of AI on jobs for cashiers: **Fewer jobs**; Expected effect of AI on jobs for factory workers: **Fewer jobs**; Expected effect of AI on jobs for medical doctors: **Fewer jobs**; Perceived AI performance compared with people at making a medical diagnosis: **AI would do this worse**; Perceived AI performance compared with people at writing a news story: **AI would do this about the same**; Perceived AI performance compared with people at making a hiring decision: **AI would do this worse**; Perceived AI performance compared with people at writing a song: **AI would do this about the same**; Perceived AI performance compared with people at deciding who gets a loan: **AI would do this about the same**; Perceived AI performance compared with people at deciding who should receive parole: **AI would do this about the same**; Perceived AI performance compared with people at driving someone from one place to another: **AI would do this about the same**; Perceived AI performance compared with people at providing customer service: **AI would do this worse**; Concern about bias in decisions made by AI: **Extremely concerned**; Concern about AI being used to impersonate people: **Extremely concerned**; Concern about personal information being misused by AI: **Extremely concerned**; Concern about inaccurate information produced by AI: **Extremely concerned**; Concern about people misunderstanding AI: **Extremely concerned**; Concern about AI reducing human connection: **Extremely concerned**; Concern about AI leading to job loss: **Extremely concerned**; Expected likelihood of AI thinking on its own: **Not too likely**; Expected likelihood of AI causing major harm: **Extremely likely**; Expected likelihood of AI making humans more productive: **Not too likely**; Expected likelihood of AI making humans happier: **Not at all likely**; Belief that AI designers consider the views of White adults: **Somewhat well**; Belief that AI designers consider the views of Black adults: **Somewhat well**; Belief that AI designers consider the views of Hispanic adults: **Somewhat well**; Belief that AI designers consider the views of Asian adults: **Somewhat well**; Belief that AI designers consider the views of men: **Somewhat well**; Expectation that AI will be trusted for important personal decisions: **No, it will not**; Confidence in the U.S. government regulating AI: **None**; Confidence in U.S. companies using AI responsibly: **None**; Self-reported exposure to information about AI: **A little**; Awareness of AI chatbots: **A little**.

Question: Thinking about the people who design artificial intelligence computer programs, how well do you think they take into account the experiences and views of women?
Options: 1. Very well; 2. Somewhat well; 3. Not too well; 4. Not at all well; 5. Not sure.
Requirements: Respond with the number only.

SocialReasoner-generated CoT

<think>The respondent reports belief that AI designers consider the views of men as "Somewhat well".
The respondent reports belief that AI designers consider the views of Hispanic adults as "Somewhat well".
The respondent reports belief that AI designers consider the views of Black adults as "Somewhat well".
The respondent reports belief that AI designers consider the views of Asian adults as "Somewhat well".
Taken together, these ordered profile details are most consistent with "Somewhat well". Therefore, the answer is 2.</think>

Prediction: 2 — Somewhat well
Observed response: 2 — Somewhat well

Analysis

The consistent “Somewhat well” ratings across men and several racial and ethnic groups suggest a generalized assessment of AI designers’ consideration of different social perspectives. The target item retains the same evaluative criterion while changing the referenced group to women, making the predicted response consistent with this cross-group assessment.

Figure 12: SocialReasoner case study on science and technology (ATP152).

Case Study: Conceptions of Democracy EVS2017

Prompt: Instruction: Answer the survey question as this respondent. Return only the option number.
Survey Context: This respondent was interviewed in 2017 in Slovakia. Interpret today and current political conditions relative to the respondent's survey year and country.
Demographic Profile: Age: 82 and older; Age group: 65+; Born in country: yes; Children count: 2; Country: Slovakia; Education level: secondary or vocational; Household income position: low; Interview language: Slovak; Marital status: widowed; Religious affiliation: yes; Religious denomination: roman catholic; Sex: female; Residential area: 20000-100000; Work status: retired/pensioned.
Perceptions: These measurements are derived from this respondent's answers to other non-target survey items. They capture the respondent's beliefs, attitudes, values, perceived risks, trust, confidence, preferences, experiences, and expectations. Generalized social trust: **Most people can be trusted**; Trust in family: **Trust completely**; Trust in people in the neighborhood: **Trust somewhat**; Trust in people known personally: **Trust completely**; Trust in people met for the first time: **Trust somewhat**; Trust in people of another religion: **Trust somewhat**; Trust in people of another nationality: **Do not trust at all**; Confidence in the church: **A great deal**; Confidence in the press: **Not very much**; Confidence in trade unions: **Quite a lot**; Confidence in the police: **A great deal**; Confidence in the civil service: **Quite a lot**; Confidence in the social security system: **A great deal**; Confidence in the European Union: **Not very much**; Confidence in the United Nations: **Quite a lot**; Confidence in the health care system: **A great deal**; Confidence in the justice system: **A great deal**; Confidence in major companies: **Not very much**; Confidence in environmental organizations: **A great deal**; Confidence in social media: **None at all**; Importance of politics in life: **Not important**; Interest in politics: **Not at all interested**; Left-right political self-placement: **9** (prompt rendering: 9 out of 10 (1 = Left, 10 = Right)); Rating of government taxing the rich and subsidizing the poor as a characteristic of democracy: **An essential characteristic of democracy**; Rating of religious authorities interpreting laws as a characteristic of democracy: **An essential characteristic of democracy**; Rating of free elections as a characteristic of democracy: **An essential characteristic of democracy**; Rating of state aid for unemployment as a characteristic of democracy: **An essential characteristic of democracy**; Rating of army takeover under incompetent government as a characteristic of democracy: **An essential characteristic of democracy**; Rating of civil rights protection against state oppression as a characteristic of democracy: **An essential characteristic of democracy**; Rating of people obeying rulers as a characteristic of democracy: **An essential characteristic of democracy**; Rating of equal rights for women and men as a characteristic of democracy: **An essential characteristic of democracy**; Importance of living in a democratically governed country: **Absolutely important**; Satisfaction with the political system: **4** (prompt rendering: 4 out of 10 (1 = Not at all satisfied, 10 = Completely satisfied)); Evaluation of a strong leader political system: **Fairly good**; Evaluation of experts making political decisions: **Very good**; Evaluation of army rule: **Fairly bad**; Evaluation of a democratic political system: **Fairly good**; Importance of religion in life: **Very important**; Religious self-identification: **A religious person**; Belief in God: **Yes**; Belief in life after death: **Yes**; Belief in hell: **Yes**; Belief in heaven: **Yes**; Belief in reincarnation: **No**; Conception of God: **A personal God**; Importance of God in life: **Very important**; Religious faith selected as an important quality for children to learn at home: **No, I would not select religious faith**.
Question: People have different views about what democracy means. Using a scale from 1 to 10, where 1 means 'not at all an essential characteristic of democracy' and 10 means 'an essential characteristic of democracy', how essential do you think the following is as a characteristic of democracy? The government makes people's incomes equal.
Options: 1. Not at all an essential characteristic of democracy; 2. 2; 3. 3; 4. 4; 5. 5; 6. 6; 7. 7; 8. 8; 9. 9; 10. An essential characteristic of democracy.
Requirements: Respond with the number only.
Additional source-profile attributes, absent from the reconstructed prompt: Employment relation: employed; Life satisfaction: 7; Political/social media frequency: never.

SocialReasoner-generated CoT

<think>The respondent rates people obeying rulers as a characteristic of democracy at an essential characteristic of democracy.
The respondent rates government taxing the rich and subsidizing the poor as a characteristic of democracy at an essential characteristic of democracy.
The respondent rates state aid for unemployment as a characteristic of democracy at an essential characteristic of democracy.
The respondent rates army takeover under incompetent government as a characteristic of democracy at an essential characteristic of democracy.
Taken together, these ordered profile details are most consistent with "An essential characteristic of democracy". Therefore, the answer is 10.</think>

Prediction: 10 — An essential characteristic of democracy
Observed response: 10 — An essential characteristic of democracy

Analysis

Classifying redistribution and unemployment assistance as essential democratic characteristics suggests that government responsibility for reducing economic disadvantage forms part of the respondent's conception of democracy. Income equalization aligns with this economic dimension of their democratic understanding, providing an interpretive basis for the response "An essential characteristic of democracy."

Figure 13: SocialReasoner case study on conceptions of democracy (EVS2017).

Case Study: Social Inequality ISSP2009

Prompt: Instruction: You are a human respondent answering a multiple-choice survey question.
Survey Context: This respondent was surveyed in Ukraine in 12.06.2009 - 21.06.2009. Interpret words such as today, nowadays, the next 12 months, five years ahead, and references to Covid-19 relative to that survey period.
Demographic Profile: Age group: 25-34; Country: Ukraine; Education level: secondary or vocational; Employment relation: employee; Ethnic identity: ukraine,Ukrainian; Household income position: lower; Household size group: small household; Marital status: married; Political orientation: no party preference; Religiosity or religious participation: religious affiliation with low participation; Sex: male; Spouse or partner's work status: part-time employed, main job; Subjective social class: middle; Union membership status: former union member; Residential area: rural; Voting status in the last election: no; Work status: employed full-time.
Perceptions: These measurements are derived from this respondent's answers to other non-target survey items. They capture the respondent's beliefs, attitudes, values, perceived risks, trust, confidence, preferences, experiences, and expectations. In deciding how much people ought to earn, how important should the amount of responsibility involved in the job be: **Very important**; In deciding how much people ought to earn, how important should the number of years spent in education and training be: **Very important**; In deciding how much people ought to earn, how important should what is needed to support a family be: **Very important**; In deciding how much people ought to earn, how important should it be whether the person has children to support: **Very important**; In deciding how much people ought to earn, how important should how well the person does the job be: **Very important**; In deciding how much people ought to earn, how important should how hard the person works at the job be: **Very important**; How important do you think coming from a wealthy family is for getting ahead in life: **Can't choose / Don't know**; How important do you think having well-educated parents is for getting ahead in life: **Can't choose / Don't know**; How important do you think having ambition is for getting ahead in life: **Can't choose / Don't know**; How important do you think hard work is for getting ahead in life: **Can't choose / Don't know**; How important do you think knowing the right people is for getting ahead in life: **Can't choose / Don't know**; How important do you think giving bribes is for getting ahead in life: **Can't choose / Don't know**; How important do you think being born a man or a woman is for getting ahead in life: **Can't choose / Don't know**; To what extent do you agree or disagree: To get all the way to the top in Ukraine today, you have to be corrupt.: **Can't choose / Don't know**; To what extent do you agree or disagree: In Ukraine, only students from the best secondary schools have a good chance to obtain a university education.: **Can't choose / Don't know**; To what extent do you agree or disagree: In Ukraine, only the rich can afford the costs of attending university.: **Can't choose / Don't know**; Is it just or unjust, right or wrong, that people with higher incomes can buy better education for their children than people with lower incomes: **Very unjust, definitely wrong**; In Ukraine, how much conflict is there between poor people and rich people: **Can't choose / Don't know**; In Ukraine, how much conflict is there between the working class and the middle class: **Can't choose / Don't know**; In Ukraine, how much conflict is there between management and workers: **Can't choose / Don't know**; In Ukraine, how much conflict is there between people at the top of society and people at the bottom: **Can't choose / Don't know**; How important do you think having a good education yourself is for getting ahead in life: **Can't choose / Don't know**; Is it just or unjust, right or wrong, that people with higher incomes can buy better health care than people with lower incomes: **Very unjust, definitely wrong**; How important do you think having political connections is for getting ahead in life: **Can't choose / Don't know**; How important do you think a person's race or ethnicity is for getting ahead in life: **Can't choose / Don't know**; To what extent do you agree or disagree: Differences in income in Ukraine are too large.: **Strongly agree**; To what extent do you agree or disagree: It is the responsibility of the government to reduce the differences in income between people with high incomes and those with low incomes.: **Strongly agree**; To what extent do you agree or disagree: The government should provide a decent standard of living for the unemployed.: **Strongly agree**; To what extent do you agree or disagree: The government should spend less on benefits for the poor.: **Strongly disagree**; How important do you think a person's religion is for getting ahead in life: **Can't choose / Don't know**.
Question: Do you think people with high incomes should pay a larger share of their income in taxes than those with low incomes, the same share, or a smaller share?
Options: 1. Much larger share; 2. Larger; 3. The same share; 4. Smaller; 5. Much smaller share; 8. Can't choose / Don't know.
Requirements: Choose the option that best suits you from the ### Options given, and respond with the number only.

SocialReasoner-generated CoT

<think>The respondent reports is it just or unjust, right or wrong, that people with higher incomes can buy better health care than people with lower incomes as "Very unjust, definitely wrong".

The respondent reports is it just or unjust, right or wrong, that people with higher incomes can buy better education for their children than people with lower incomes as "Very unjust, definitely wrong".

The respondent reports how important do you think hard work is for getting ahead in life as "Can't choose / Don't know".

The respondent reports how important do you think having ambition is for getting ahead in life as "Can't choose / Don't know".

Taken together, these ordered profile details are most consistent with "Much larger share". Therefore, the answer is 1.</think>

Analysis

Strong objections to income-based advantages in education and health care suggest a concern with distributive fairness, particularly the extent to which economic resources determine access to better services. Assigning high earners a greater proportional tax burden is consistent with this concern, supporting the predicted preference for a “Much larger share.”

Prediction: 1 — Much larger share
Observed response: 1 — Much larger share

Figure 14: SocialReasoner case study on social inequality (ISSP2009).

Case Study: Work Orientations ISSP2015

Prompt: Instruction: You are a human respondent answering a multiple-choice survey question.
Survey Context: This respondent was surveyed in Austria in 09.07.2015 - 16.08.2016. Interpret words such as today, nowadays, the next 12 months, five years ahead, and references to Covid-19 relative to that survey period.
Demographic Profile: Age group: 55-64; Number of children in household: no children; Country: Austria; Education level: secondary or vocational; Employment relation: employee; Household income position: lower; Household size group: one-person household; Marital status: divorced from spouse/ legally separated from civil partner; Political orientation: center; Religiosity or religious participation: religious affiliation with low participation; Sex: female; Steady partner status: no partner; Subjective social class: middle; Union membership status: not a union member; Residential area: rural village; Voting status in the last election: yes; Work status: retired.
Perceptions: These measurements are derived from this respondent's answers to other non-target survey items. They capture the respondent's beliefs, attitudes, values, perceived risks, trust, confidence, preferences, experiences, and expectations. How important is high income to you personally: **Not important**; How important are opportunities for advancement to you personally: **Important**; How important is having an interesting job to you personally: **Very important**; How important is being able to work independently to you personally: **Very important**; How important is having a job that is useful to society to you personally: **Very important**; How important is having personal contact with other people at work to you personally: **Very important**; To what extent do you agree or disagree that strong trade unions are bad for the economy: **Neither agree nor disagree**; Would you like to have a paid job now or in the future: **No**; To what extent do you worry about losing your main source of economic support: **I worry a little**; To what extent do you agree or disagree that senior citizens who continue working take jobs away from younger people: **Strongly disagree**.
Question: To what extent do you agree or disagree that a job is just a way of earning money?
Options: 1. Strongly agree; 2. Agree; 3. Neither agree nor disagree; 4. Disagree; 5. Strongly disagree; 8. Can't choose.
Requirements: Choose the option that best suits you from the ### Options given, and respond with the number only.

SocialReasoner-generated CoT

<think>The respondent reports how important is high income to you personally as "Not important".

The respondent reports to what extent do you agree or disagree that senior citizens who continue working take jobs away from younger people as "Strongly disagree".

The respondent reports how important is having an interesting job to you personally as "Very important".

The respondent reports how important is having personal contact with other people at work to you personally as "Very important".

Taken together, these ordered profile details are most consistent with "Strongly disagree". Therefore, the answer is 5.</think>

Prediction: 5 — Strongly disagree
Observed response: 5 — Strongly disagree

Analysis

The respondent prioritizes interesting work and workplace interaction over high income, suggesting that employment is valued for its intrinsic and relational benefits rather than financial returns alone. Describing a job as “just a way of earning money” excludes these non-monetary benefits, directly conflicting with this conception of work and supporting the predicted strong disagreement.

Figure 15: SocialReasoner case study on work orientations (ISSP2015).

Case Study: Role of Government ISSP2016

Prompt (excerpt): Instruction: You are a human respondent answering a multiple-choice survey question.
Survey Context: This respondent was surveyed in South Korea in 27.06.2016 - 07.10.2016. Interpret words such as today, nowadays, the next 12 months, five years ahead, and references to Covid-19 relative to that survey period.
Demographic Profile: Age group: 45-54; Number of children in household: no children; Country: South Korea; Education level: primary or basic; Employment relation: self-employed without employees; Household income position: top 20% within the country; Household size group: single-person household; Marital status: single or never married; Religiosity or religious participation: no religious affiliation; Sex: male; Steady partner status: no partner; Subjective social class: lower; Residential area: rural; Voting status in the last election: yes; Work status: in paid work.
Perceptions: These measurements are derived from this respondent's answers to other non-target survey items. They capture the respondent's beliefs, attitudes, values, perceived risks, trust, confidence, preferences, experiences, and expectations. In general, should people obey the law without exception, or are there exceptional occasions on which people should follow their consciences even if it means breaking the law: **Obey the law without exception**; Generally, how would you describe taxes in South Korea today for those with high incomes: **Much too low**; Generally, how would you describe taxes in South Korea today for those with low incomes: **Much too high**; In general, how often do you think tax authorities in South Korea make sure people pay their taxes: **Often**; In general, how often do you think major private companies in South Korea try to avoid paying their taxes: **Often**; Which type of justice error do you think is worse: to convict an innocent person or to let a guilty person go free: **Let a guilty person go free**; Who should primarily provide care for older people in South Korea: **Family, relatives or friends**; To what extent do you agree or disagree: People like me don't have any say about what the government does: **Strongly agree**; To what extent do you agree or disagree: I feel that I have a pretty good understanding of the important political issues facing our country: **Agree**; To what extent do you agree or disagree: People we elect as MPs try to keep the promises they have made during the election: **Neither agree nor disagree**; Should it be the government's responsibility to provide health care for the sick: **Definitely should not be**; Should the government of South Korea have the right to keep people under video surveillance in public areas: **Definitely should have right**; Should the government of South Korea have the right to monitor e-mails and other information exchanged on the Internet: **Definitely should have right**; [...]; How interested would you say you personally are in politics: **Somewhat interested**; In your opinion, about how many politicians in South Korea are involved in corruption: **Almost all**; Should organising public meetings to protest against the government be allowed: **Definitely allowed**; Should people who want to overthrow the government by revolution be allowed to hold public meetings to express their views: **Probably allowed**; Should people who want to overthrow the government by revolution be allowed to publish books expressing their views: **Definitely allowed**; Are you in favour of or against cuts in government spending: **Against**; Would you like to see more or less government spending on the environment? Remember that spending more might require a tax increase.: **Spend much more**; Would you like to see more or less government spending on education? Remember that spending more might require a tax increase.: **Spend much more**; Would you like to see more or less government spending on the military and defence? Remember that spending more might require a tax increase.: **Spend much more**; Would you like to see more or less government spending on old age pensions? Remember that spending more might require a tax increase.: **Spend much more**; Would you like to see more or less government spending on unemployment benefits? Remember that spending more might require a tax increase.: **Spend much more**; Would you like to see more or less government spending on culture and the arts? Remember that spending more might require a tax increase.: **Spend more**; Should it be the government's responsibility to provide a job for everyone who wants one: **Definitely should not be**; Should it be the government's responsibility to keep prices under control: **Probably should not be**; Should it be the government's responsibility to provide a decent standard of living for the old: **Definitely should be**; Should it be the government's responsibility to reduce income differences between the rich and the poor: **Definitely should not be**; Should it be the government's responsibility to provide decent housing for those who cannot afford it: **Definitely should be**; Which person or organisation has the most influence on the actions of the government of South Korea: **Organised crime**.

Question: Should it be the government's responsibility to give financial help to university students from low-income families?
Options: 1. Definitely should be; 2. Probably should be; 3. Probably should not be; 4. Definitely should not be; 8. Can't choose / Don't know.
Requirements: Choose the option that best suits you from the ### Options given, and respond with the number only.

SocialReasoner-generated CoT

<think>The respondent reports should it be the government's responsibility to provide decent housing for those who cannot afford it as "Definitely should be".

The respondent reports would you like to see more or less government spending on education? Remember that spending more might require a tax increase as "Spend much more".

The respondent reports should it be the government's responsibility to provide a decent standard of living for the old as "Definitely should be".

The respondent reports would you like to see more or less government spending on the military and defence? Remember that spending more might require a tax increase as "Spend much more".

Taken together, these ordered profile details are most consistent with "Definitely should be". Therefore, the answer is 1.</think>

Analysis

Support for housing assistance and substantially increased education spending suggests two relevant commitments: assistance for those with limited means and public investment in education. Financial aid for university students from low-income families combines these commitments, providing convergent grounds for the predicted endorsement of government responsibility.

Prediction: 1 — Definitely should be
Observed response: 1 — Definitely should be

Figure 16: SocialReasoner case study on the role of government (ISSP2016).

Case Study: Social Inequality ISSP2019

Prompt: Instruction: You are a human respondent answering a multiple-choice survey question.
Survey Context: This respondent was surveyed in Chile in 26.04.2019 - 03.06.2019. Interpret words such as today, nowadays, the next 12 months, five years ahead, and references to Covid-19 relative to that survey period.
Demographic Profile: Age group: 65+; Number of children in household: one child [school entry age] to 17; Country: Chile; Education level: primary or basic; Employment relation: employee; Household size group: small household; Marital status: widowed; Religiosity or religious participation: low religious participation; Sex: female; Steady partner status: no partner; Subjective social class: middle; Union membership status: not a union member; Residential area: urban; Voted last election: no; Work status: retired.
Perceptions: These measurements are derived from this respondent's answers to other non-target survey items. They capture the respondent's beliefs, attitudes, values, perceived risks, trust, confidence, preferences, experiences, and expectations. In your opinion, in Chile, how much conflict is there between poor people and rich people: **Strong conflicts**; In your opinion, in Chile, how much conflict is there between management and workers: **Strong conflicts**; In your opinion, in Chile, how much conflict is there between people born in Chile and people from other countries who have come to live in Chile: **Very strong conflicts**; Thinking ahead 10 years from now, where do you think you will be on a scale from 1 at the bottom to 10 at the top: **5**; In deciding how much people ought to earn, how important should the amount of responsibility involved in the job be: **Very important**; In deciding how much people ought to earn, how important should the number of years spent in education and training be: **Very important**; In deciding how much people ought to earn, how important should it be whether the person has children to support: **Very important**; Which type of society best describes Chile today: **B: Pyramid with small elite at top, more people in middle, most at bottom.**; How important do you think coming from a wealthy family is for getting ahead in life: **Fairly important**; How important do you think hard work is for getting ahead in life: **Very important**; How important do you think knowing the right people is for getting ahead in life: **Fairly important**; How important do you think giving bribes is for getting ahead in life: **Not important at all**; How important do you think being born a man or a woman is for getting ahead in life: **Not important at all**; Generally, how would you describe taxes in Chile today for those with high incomes: **Too low**; Generally speaking, would you say that people can be trusted or that you can't be too careful in dealing with people: **You almost always can't be too careful in dealing with people**; How important do you think having a good education yourself is for getting ahead in life: **Very important**; Is it just or unjust, right or wrong, that people with higher incomes can buy better health care than people with lower incomes: **Very unjust, definitely wrong**; To what extent do you agree or disagree with the statement: People from poor countries should be allowed to work in wealthy countries: **Strongly agree**; How important do you think having political connections is for getting ahead in life: **Not very important**; How important do you think a person's race or ethnicity is for getting ahead in life: **Not important at all**; To what extent do you agree or disagree with the statement: Differences in income in Chile are too large: **Agree**; How successful do you think the government in Chile is nowadays in reducing the differences in income between people with high incomes and people with low incomes: **Very unsuccessful**; Is it just or unjust, right or wrong, that people with higher incomes can buy better education for their children than people with lower incomes: **Very unjust, definitely wrong**; How important do you think a person's religion is for getting ahead in life: **Not important at all**; To what extent do you agree or disagree with the statement: It is the responsibility of the government to reduce the differences in income between people with high incomes and those with low incomes: **Strongly agree**; To what extent do you agree or disagree with the statement: It is the responsibility of private companies to reduce the differences in pay between their employees with high pay and those with low pay: **Agree**.
Question: How angry do you feel when you think about differences in wealth between the rich and the poor in Chile?
Options: 0. Not angry at all; 1. 1; 2. 2; 3. 3; 4. 4; 5. 5; 6. 6; 7. 7; 8. 8; 9. 9; 10. Extremely angry.
Requirements: Choose the option that best suits you from the ### Options given, and respond with the number only.

SocialReasoner-generated CoT

<think>The respondent reports is it just or unjust, right or wrong, that people with higher incomes can buy better education for their children than people with lower incomes as "Very unjust, definitely wrong".

The respondent reports is it just or unjust, right or wrong, that people with higher incomes can buy better health care than people with lower incomes as "Very unjust, definitely wrong".

The respondent reports how successful do you think the government in Chile is nowadays in reducing the differences in income between people with high incomes and people with low incomes as "Very unsuccessful".

The respondent reports in your opinion, in Chile, how much conflict is there between people born in Chile and people from other countries who have come to live in Chile as "Very strong conflicts".

Taken together, these ordered profile details are most consistent with "Extremely angry". Therefore, the answer is 10.</think>

Prediction: 10 — Extremely angry
Observed response: 10 — Extremely angry

Analysis

The highlighted evidence combines strong objections to income-based advantages in education and health care with perceived government failure to reduce income differences. Together, these judgments suggest that inequality is experienced as both unjust and insufficiently addressed, linking the respondent's fairness concerns and dissatisfaction with corrective efforts to the predicted extreme anger.

Figure 17: SocialReasoner case study on social inequality (ISSP2019).

Case Study: Health & Health Care ISSP2021

Prompt: Instruction: You are a human respondent answering a multiple-choice survey question.
Survey Context: This respondent was surveyed in Australia in 17.05.2021 - 01.05.2022. Interpret words such as today, nowadays, the next 12 months, five years ahead, and references to Covid-19 relative to that survey period.
Demographic Profile: Age group: 45-54; Country: Australia; Education level: some college; Employment relation: employee; Household size group: two-person household; Marital status: married; Political orientation: right; Religiosity or religious participation: religious affiliation with low participation; Sex: female; Spouse partner work status: retired; Steady partner status: yes, have partner; live in same household; Subjective social class: upper; Union membership status: current union member; Residential area: town or small city; Voting status in the last election: yes; Work status: in paid work.
Perceptions: These measurements are derived from this respondent's answers to other non-target survey items. They capture the respondent's beliefs, attitudes, values, perceived risks, trust, confidence, preferences, experiences, and expectations. To what extent do you agree or disagree that overall, vaccinations do more harm than good: **Strongly disagree**; To what extent do you agree or disagree that getting immunity by becoming ill is better than having a vaccination: **Strongly disagree**; How much confidence do you have in the health care system in your country: **A great deal of confidence**; To what extent do you agree or disagree that doctors' medical skills are not as good as they should be: **Disagree**; To what extent do you agree or disagree that doctors care more about their earnings than about their patients: **Disagree**; In general, how fair or unfair do you think the health care system is in your country: **Somewhat fair**; To what extent do you agree or disagree that people use health care services more than is necessary: **Neither agree nor disagree**; To what extent do you agree or disagree that the government should provide only limited health care services: **Disagree**; To what extent do you agree or disagree that the health care system in your country is inefficient: **Disagree**; How willing would you be to pay higher taxes to improve the health care system in your country: **Fairly unwilling**; Should people have access to publicly funded health care even if they are not citizens of your country: **Disagree**; Compared with poor people, is access to health care easier or harder for rich people: **Somewhat easier**; Compared with men, is access to health care easier or harder for women: **About the same**; To what extent do you agree or disagree that health problems are caused by genes: **Neither agree nor disagree**; To what extent do you agree or disagree that health problems are caused by poverty: **Disagree**; To what extent do you agree or disagree that the internet is useful for checking a doctor's advice: **Neither agree nor disagree**; To what extent do you agree or disagree that most people become very overweight because they are lazy: **Disagree**; To what extent do you agree or disagree that most people contracted Covid-19 because they were careless: **Disagree**; During a severe epidemic, should the government have the right to shut down businesses and public places: **Definitely should have the right**; During a severe epidemic, should the government have the right to demand that people stay at home: **Probably should have the right**; During a severe epidemic, should the government have the right to use digital surveillance to track infected people: **Probably should have the right**; During a severe epidemic, should the government have the right to suspend compulsory education: **Definitely should have the right**.
Question: To what extent do you agree or disagree that doctors can be trusted?
Options: 1. Strongly agree; 2. Agree; 3. Neither agree nor disagree; 4. Disagree; 5. Strongly disagree.
Requirements: Choose the option that best suits you from the ### Options given, and respond with the number only.

SocialReasoner-generated CoT

<think>The respondent reports to what extent do you agree or disagree that doctors care more about their earnings than about their patients as "Disagree".
The respondent reports to what extent do you agree or disagree that doctors' medical skills are not as good as they should be as "Disagree".
The respondent reports how much confidence do you have in the health care system in your country as "A great deal of confidence".
The respondent reports to what extent do you agree or disagree that the health care system in your country is inefficient as "Disagree".
Taken together, these ordered profile details are most consistent with "Agree". Therefore, the answer is 2.</think>

Analysis

Rejecting claims of inadequate medical skills and the prioritization of earnings over patient care suggests confidence in doctors' competence and professional motives. Reinforced by strong confidence in the health care system, these evaluations provide complementary professional and institutional grounds for the predicted agreement that doctors can be trusted.

Prediction: 2 — Agree
Observed response: 2 — Agree

Figure 18: SocialReasoner case study on health and health care (ISSP2021).