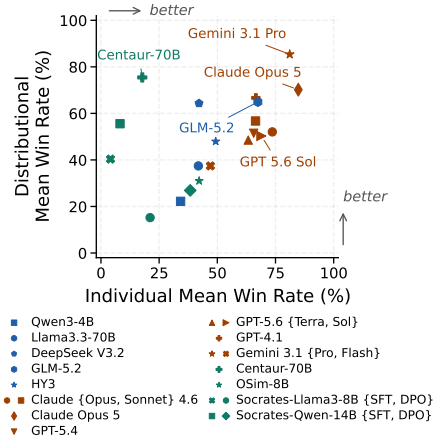
BEHAVIORBENCH: BENCHMARKING FOUNDATION MODELS FOR BEHAVIORAL SCIENCE TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models have been increasingly applied to behavioral science domains such as psychology, sociology, and economics. While these models show promise in tasks such as survey response prediction and human-subject experiment simulation, there remains no systematic understanding of how well they perform across diverse behavioral science tasks. We introduce `BehaviorBench`, a comprehensive benchmark that evaluates foundation models along four core capabilities: (1) behavior prediction and simulation, (2) strategic decision-making, (3) subject-trait inference, and (4) behavioral knowledge application. Crucially, `BehaviorBench` evaluates model outputs at both the individual and distributional levels, capturing not only per-subject accuracy but also population-level alignment, an essential requirement for behavioral validity. Our evaluation shows that `BehaviorBench` remains challenging for leading general-purpose LLMs and behavior foundation models that are specifically trained with behavioral data. We find that individual-level and distributional performance do not always align. General-purpose LLMs tend to underestimate the diversity of human responses, whereas behavior foundation models often lag behind at individual-level prediction. Our investigation further demonstrates how fine-tuning on diverse behavioral data can improve both individual-level prediction and distributional alignment, balancing these two objectives. Our results highlight the importance of evaluation at both individual and distributional levels, establishing `BehaviorBench` as a foundation for developing and assessing behaviorally aligned AI systems.



BehaviorBench Leaderboards

Individual Level

#	Model	Mean WR
1	Claude Opus 5	84.8%
2	Gemini 3.1 Pro	81.1%
3	Claude Opus 4.6	73.6%
4	GPT 5.6 Sol	68.9%
5	GLM-5.2	67.4%
6	GPT 4.1	66.6%
7	Claude Sonnet 4.6	66.4%
8	GPT 5.4	65.6%
9	GPT 5.6 Terra	63.3%
10	HY3	49.3%
11	Gemini 3.1 Flash	47.0%
12	OSim-8B	42.2%
…	…	…

Distributional Level

#	Model	Mean WR
1	Gemini 3.1 Pro	85.4%
2	Centaur-70B	75.4%
3	Claude Opus 5	70.2%
4	GPT 4.1	66.7%
5	GLM-5.2	64.9%
6	DeepSeek V3.2	64.3%
7	Claude Sonnet 4.6	56.7%
8	Socrates-14B-SFT	55.6%
9	Claude Opus 4.6	52.0%
10	GPT 5.4	51.5%
11	GPT 5.6 Sol	50.3%
12	GPT 5.6 Terra	48.5%
…	…	…

Open-source LLMs Proprietary LLMs Behavioral foundation models

Figure 1: Aggregated evaluation results of foundation models on `BehaviorBench`. *Left:* individual- vs. distributional-level performance in mean win rate (%) across 20 models. *Right:* Snapshots of the top 12 models on the BehaviorBench leaderboards, with models ranked at the individual and distributional levels by mean win rate (Mean WR, %).

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.

1 INTRODUCTION

Foundation models have transformed many fields. Trained on vast and diverse data, they can adapt across tasks and are increasingly used not only for prediction, but also for reasoning, simulation, and scientific discovery (Bommasani et al., 2021; Park et al., 2023; Wang et al., 2023). One field with particularly transformative potential is *behavioral science*: the science of understanding, predicting, and shaping human behaviors. Drawing on multiple disciplines, including psychology, sociology, anthropology, and economics, behavioral science provides foundations for understanding decision-making dynamics at individual, group, and societal levels (Gintis, 2014).

Recent work has started to apply general-purpose foundation models, including commercial and open-source large language models, to behavioral science (Demszky et al., 2023; Meng, 2024; Bail, 2024). These models show promise in improving survey research (Argyle et al., 2023; Grossmann et al., 2023), supporting agent-based modeling (Park et al., 2023; 2024), and predicting human-subject experiment results (Hewitt et al., 2024; Lippert et al., 2024). In parallel, researchers are also developing behavioral-science-specific foundation models by fine-tuning LLMs on behavior-related data. These models have demonstrated improved performance in simulating human behavior, predicting survey responses, and supporting behavioral research workflows (Xie et al., 2025; Binz et al., 2024; Kolluri et al., 2025).

Despite this growing body of work, we still lack a systematic understanding of how well foundation models perform across diverse behavioral science tasks, contexts, and populations. Existing behavior benchmarks are often limited by a narrow task family, such as survey response prediction or behavior simulation (Zhou et al., 2026; Hu et al., 2025; Suh et al., 2025; Wang et al., 2025). As a result, models' capabilities in important behavior science tasks, such as inferring latent traits of human subjects or applying behavioral knowledge to scientific research, remain unmeasured.

To address this limitation, we introduce `BehaviorBench`, a comprehensive benchmark for systematically evaluating foundation models on behavioral science tasks. We organize our evaluation around the observation that human behavior is shaped by multiple factors like the behavioral context, the subject's traits, and underlying motivations (Xie et al., 2025). Formally, human behavior can be represented as a conditional probability

$$p(y \mid x, c; \mathbb{K}), \tag{1}$$

where y denotes a behavior, x the characteristics of the human subject, c the contextual variables defining the environment or intervention, and $\mathbb{K}$ represents relevant behavioral knowledge. With the majority of existing benchmarks focusing only on the prediction of y, however, to best support behavioral science, a behavior foundation model should support multiple types of inference within this framework. As a consequence, `BehaviorBench` evaluates models along four core capabilities: (1) behavior prediction and simulation (forward inference over y to mimic a human subject), (2) strategic decision-making (the generation of y to achieve a specified goal), (3) subject-trait inference (inverse inference over x), and (4) the application of behavioral science knowledge ($\mathbb{K}$) to research problems. Built on a diverse collection of datasets, `BehaviorBench` comprises 12 distinct tasks that span these four capabilities.

In addition, for central behavioral tasks such as behavior simulation, researchers care about the accuracy at two complementary levels: *individual level*, where the objective is to predict individual outcomes; and *distributional level*, where the objective is to predict the outcome distribution of the population (Xie et al., 2026). However, most existing benchmarks emphasize only one of these levels (Zhou et al., 2026; Hu et al., 2025; Wang et al., 2025) (Table 2). In contrast, our `BehaviorBench` provides dedicated evaluations at both the *individual* and the *distributional* levels.

Our main contributions are:

- We introduce `BehaviorBench`, a comprehensive benchmark covering 12 behavioral science tasks across four capabilities, with evaluation at both the individual and distributional levels.
- We systematically evaluate 20 LLMs, including general-purpose frontier LLMs, and behavior foundation models that are specifically trained on behavioral data. Our results show that `BehaviorBench` is challenging for all models evaluated, with Gemini 3.1 Pro and Claude Opus 5 standing out on the leaderboards.

- We find that individual-level and distributional performance do not always align. In particular, general-purpose LLMs tend to underestimate the diversity of human responses, despite more accurately predicting the population mean, whereas behavior foundation models better preserve response diversity.
- Our further investigation into fine-tuning LLMs with diverse behavioral data shows the possibility of simultaneously improving the individual-level and distributional performance, demonstrating the pathway to behaviorally aligned AI systems.

2 BENCHMARK CONSTRUCTION

2.1 DATA

Capturing the breadth of human behavior demands data of comparable breadth. Following Xie et al. (2025), `BehaviorBench` draws on three sources of behavior data to construct the evaluation. Details are in Appendix B.

1. **Experimental data.** Experimental records capture behavioral patterns of diverse subjects in controlled environments, offering key insights into individual and group decision-making. We consider a collection of experiments from MobLab[1], which covers a diverse population of human subjects and their play records in multiple economics games (Mei et al., 2024). We choose this dataset because it is one of the largest collections of controlled economic-game records, covering tens of thousands of subjects from more than 50 countries. `BehaviorBench` uses an evaluation subsample of 12,685 behavioral observations from 6,279 unique subjects.
2. **Survey data.** Survey data provide self-reported insights into personality traits, attitudes, and preferences. We use the publicly available Big Five Personality Test dataset[2] as a data source. This dataset consists of demographic information of subjects (e.g., race, age, gender), along with their answers to 50 survey questions. The answers can be aggregated into five personality scores. We choose this dataset because it is one of the largest publicly available personality-survey datasets, with around one million responses and respondents from 158 countries. `BehaviorBench` includes an evaluation subsample of 6,000 responses from 1,886 unique subjects.
3. **Literature data.** The behavioral science literature documents existing theories and scientific findings about human behavior. We collect titles and abstracts of articles from *American Economic Review* and *Nature Human Behaviour* that were published in 2025. We then follow Zhang et al. (2025) to decompose each title–abstract pair into a structured research workflow that reflects how the studies are done. Such data are included in `BehaviorBench` to evaluate foundation models' ability in scientific reasoning and assisting behavioral science research.

2.2 EVALUATION TASKS

By formulating the generation of human behavior as $p(y \mid x, c, \mathbb{K})$, `BehaviorBench` evaluates foundation models with four categories of behavioral tasks (Table 1): (1) behavior prediction and simulation (forward inference over y), (2) strategic decision-making (the generation of y to achieve a goal), (3) subject-trait inference (inverse inference over x), and (4) the application of behavioral science knowledge ($\mathbb{K}$) for scientific reasoning. Formally, let $\mathcal{Y}$ denote the space of human behaviors (e.g., actions in experiments and survey responses), $\mathcal{X}$ the space of human subject traits (e.g., demographics and personalities), and $\mathcal{C}$ the space of contextual variables (e.g., experiment setup and intervention).

2.2.1 BEHAVIOR PREDICTION AND SIMULATION

Predicting and simulating human behavior is at the core of behavior modeling (Wang et al., 2025; Binz et al., 2024; Kolluri et al., 2025). Behavior prediction arises naturally at two levels: the individual level and the distributional level.

[1] MobLab: `https://www.moblab.com/`.

[2] `https://www.kaggle.com/datasets/lucasgreenwell/ocean-five-factor-personality-test-responses`.

Capability	Task	# Eval. Data	Eval. Metric	
			Ind.	Dist.
Behavior prediction and simulation	Single-round game behavior simulation (Game Behav. Sim.)	200/game	–	W
	Multi-round game behavior prediction (Multi-Round Pred.)	500/game	MAE	W
	Single-round game behavior prediction given observations from other games (Across-Ctx Pred.)	750/game	MAE	W
	Survey response prediction given demographics (Demo. To Resp.)	1,000	Acc	W
	Sequential survey response prediction (Seq. Resp. Pred.)	1,000	Acc	W
	Masked survey response prediction (Masked Resp. Pred.)	1,000	Acc	W
Strategic decision-making	Strategic game play	1,000	Win (%)	–
Subject-trait inference	Personality score prediction given demographics (Demo. To Pers.)	1,000	MAE	W
	Personality score prediction given scores from other dimensions (Across-Dim Pers. Pred.)	1,000	MAE	W
	Age prediction given personality scores (Pers. To Demo.)	1,000	MAE	W
Knowledge application	Scientific workflow prediction	1,084	BLEURT	–
	Economics contest problem solving	124	Acc	–

Table 1: Evaluation tasks in `BehaviorBench`, by core capabilities. **#Eval. Data**: evaluation set sizes. Metrics: individual level (**Ind.**) and distributional level (**Dist.**, "W" stands for Wasserstein distance). Per-task data sources and processing procedures are detailed in Appendix B.

At the individual level, the task is to predict the behavior of a particular subject in a given context. Formally, this can be represented as a projection

$$\mathcal{X} \times \mathcal{Y} \times \mathcal{C} \to \mathcal{Y}. \tag{2}$$

Concretely, a model predicts a subject's behavior y (e.g., their choice in the next round) from the subject's characteristics x, the context c (e.g., experiment instructions and setup), and partial information about the subject's behavior y', which can be empty.

At the distributional level, the target is to predict a behavior distribution across a population. To obtain a behavior distribution Y, we run individual-level predictions and then evaluate whether Y aligns with the empirical human behavior distribution. `BehaviorBench` constructs the evaluation of this capability from two data sources below.

Game play simulation and prediction. We define three simulation/prediction tasks. (1) *Single-round game behavior simulation.* Given the game instructions, a model predicts the distribution of the population's actions in a single (first) round. The purpose is to measure whether a model distributionally aligns with the human population when playing a game. (2) *Multi-round game behavior prediction.* Given the game instructions and a subject's behavior in prior rounds, a model predicts their action in the next round. Different from (1), here we condition on a subject's decision trajectory. (3) *Across-context prediction.* Given a subject's behavior in other games, a model predicts the first-round behavior in a target game.

Survey response prediction. Using the Big Five responses described in Section 2.1, we define three prediction tasks. (1) *Survey response prediction given demographics.* Given a subject's demographics and a single survey item, a model predicts the subject's response. This task is consistent with the setting in prior works (Santurkar et al., 2023; Kolluri et al., 2025). (2) *Sequential survey response prediction.* Given a subject's responses to prior items in survey order, a model sequentially predicts the response to the next item. (3) *Masked survey response prediction.* Given all other responses, a model predicts the response to one masked survey item.

2.2.2 Strategic Decision-Making

Section 2.2.1 concerns predicting and simulating a subject's behavior. Here, we examine a distinct task: making strategic decisions to achieve a specified goal. In this setting, actions are taken on one's

own behalf, rather than mimicking a particular human subject. This capability is especially important as AI and LLM agents increasingly make autonomous decisions and interact with humans (Mei et al., 2024; Jackson et al., 2025). Formally, this task requires a model to make the projection

$$\mathcal{C} \times \mathcal{Y} \to \mathcal{Y}. \tag{3}$$

Concretely, a model generates its own decision y toward the specified goal from the context c (e.g., experiment instructions) and other subjects' actions. Prior work has used economic games to measure this capability (Shapira et al., 2024; Guo et al., 2024; Huang et al., 2025; Duan et al., 2024). However, most prior works have LLMs play against other LLMs; few have LLMs play against human player records (Shapira et al., 2024).

`BehaviorBench` uses human play records of the Beauty Contest game on MobLab to construct an evaluation. In Beauty Contest, each player picks a number in $[0, 100]$ for every round. The winner is whoever comes closest to $\frac{2}{3}$ of the group average. At the end of each round, each player gets feedback on whether they won that round. This game involves opponent modeling and extensive reasoning (Nagel, 1995; Ho et al., 1998).

2.2.3 Subject-trait Inference

In Sections 2.2.1 and 2.2.2, we introduced tasks for predicting, simulating, and generating behavior (y) under various settings. Here, we introduce the inverse tasks for inferring subjects' traits (x). Characterizing individual subjects and populations is a core goal of behavioral science (Yang & Zhai, 2022; Rychalska et al., 2023; Demszky et al., 2023; Bleidorn et al., 2019). At the individual level, the task is to infer a latent characteristic of a particular subject from their observed behavior. Formally, a foundation model is tasked with the projection

$$\mathcal{Y} \times \mathcal{X} \times \mathcal{C} \to \mathcal{X}. \tag{4}$$

At the distributional level, we run individual-level inferences over a population and then evaluate whether the predicted distribution aligns with the empirical human trait distribution.

We introduce three tasks. (1) *Personality score prediction given demographics.* Given a subject's demographics, a model predicts their score on a Big Five personality dimension. (2) *Across-dimension personality score prediction.* Given a subject's scores on four personality dimensions, a model predicts their score on the remaining dimension. (3) *Age prediction given personality scores.* Given a subject's five-dimensional personality, a model predicts the subject's age.

2.2.4 Behavioral Knowledge Application

Beyond prediction and simulation, a behavior foundation model should be able to represent and apply behavioral science knowledge. These tasks take the form of applying knowledge $\mathbb{K}$ to solve the input question q.

Research workflow reasoning. Following Zhang et al. (2025), we decompose a paper's title and abstract into five research-workflow aspects: context, key idea, method, outcome, and projected impact. We then define five generation tasks: key idea generation, method recommendation, outcome prediction, impact prediction, and title prediction. This group of tasks reflects models' ability to assist behavioral research workflows.

Complex problem solving. To test whether the model can apply domain knowledge to unseen problems, we use multiple-choice questions from the International Economics Olympiad (IEO) contests, spanning from 2018 to 2025[3]. Model choices are compared with official solutions.

2.3 Metrics

For individual-level evaluation, we use:

- **Mean absolute error (MAE)** for the prediction of numeric quantities, including game actions, trait scores, and age. We choose MAE because it preserves the ordinal structure of the response.
- **Accuracy** for categorical predictions, including predicting survey responses and answering IEO questions (Kolluri et al., 2025).

[3]International Economics Olympiad (IEO) official website: `https://ieo-official.org/`.

Capability	Predict & Sim. Behav.	Strategic Decision-Making	Infer Subj. Traits	Apply Knowledge	Eval. Level		Public Benchmark
Targeted Variable		y	x	$\mathbb{K}$	Indiv.	Distrib.	
BehaviorBench (Ours)	✓	✓	✓	✓	✓	✓	✓
SOUL-Index (Zhou et al., 2026)	✓				✓		✓
SocioBench (Wang et al., 2025)	✓				✓		✓
SubPOP (Suh et al., 2025)	✓					✓	
SimBench (Hu et al., 2025)	✓					✓	✓
GLEE (Shapira et al., 2024)		✓			✓		✓
GTBench (Duan et al., 2024)		✓			✓		✓
GAMA-Bench (Huang et al., 2025)		✓			✓		✓

Table 2: Comparison of BehaviorBench with existing behavioral benchmarks. BehaviorBench is the only one evaluating all four capabilities at both individual and distributional levels.

- **Win rate** for strategic decision-making (i.e., the Beauty Contest game), defined as the fraction of sessions that a model wins over human players.
- **BLEURT** (Sellam et al., 2020) for open-ended generation in the research workflow task, following Zhang et al. (2025).

For distribution-level evaluation, we use the **Wasserstein distance** (noted as W) to compare the predicted behavior distribution against the observed human distribution. Following Suh et al. (2025); Kolluri et al. (2025), we choose W because it captures both the shape and the mean of two distributions. More details on the evaluation metrics are in Appendix C.

2.4 Comparison with Existing Benchmarks

Table 2 compares BehaviorBench with existing benchmarks for behavioral-science-related tasks. Existing benchmarks primarily target narrow task families: SOUL-Index (Zhou et al., 2026), SocioBench (Wang et al., 2025), SubPOP (Suh et al., 2025), and SimBench (Hu et al., 2025) evaluate behavior prediction from demographics or context, while GLEE (Shapira et al., 2024), GTBench (Duan et al., 2024), and GAMA-Bench (Huang et al., 2025) evaluate strategic decision-making in games. These benchmarks focus on behavior y. **None of them evaluates the inference of subject traits** x **or the application of behavioral knowledge** $\mathbb{K}$. Moreover, BehaviorBench supports evaluation at both the individual and distributional levels, while previous works tend to only evaluate at one level. We also identify several publicly available datasets that could potentially be developed into benchmarks, including OpinionQA (Santurkar et al., 2023), Psych-101 (Binz et al., 2024), and SocSci210 (Kolluri et al., 2025). However, these datasets likewise do not fully cover the four capability categories.

3 Benchmarking Foundation Models

3.1 Models to Evaluate

We benchmark three types of foundation models representing different model types: (1) open-source general-purpose LLMs, (2) proprietary general-purpose LLMs, and (3) behavior foundation models that are trained on behavior-related data.

Open-source LLMs. We include open-source LLMs with various model sizes, including Qwen3-4B (Team, 2025), Llama-3.3-70B (Team, 2024), HY3 (Tencent Hy Team, 2026) (295B), DeepSeek V3.2 (DeepSeek-AI, 2025) (671B), and GLM-5.2 (GLM, 2026) (753B), spanning from small dense models to bigger mixture-of-experts models.

Proprietary LLMs. We include three families of widely used frontier proprietary models. Within each family we include different capability tiers. For Claude, we include Sonnet 4.6 (Anthropic, 2026c), Opus 4.6 (Anthropic, 2026a) and Opus 5 (Anthropic, 2026b). For GPT, we include leading reasoning models, GPT-5.4 (OpenAI, 2026a), GPT 5.6 Terra and Sol (OpenAI, 2026b), together with GPT-4.1 (OpenAI, 2025a) as the strongest non-reasoning GPT model. For Gemini, we include Gemini 3.1 Flash and Gemini 3.1 Pro (Google, 2026). We use the default inference hyperparameters for all models.

Capability	Behavior Prediction & Simulation											Strategic
Task	Game Behav. Sim.	Multi-Round Pred.		Across-Ctx Pred.		Demo. To Resp.		Seq. Resp. Pred.		Masked Resp. Pred.		Game Play
Metric	W ↓	MAE ↓	W ↓	MAE ↓	W ↓	Acc ↑	W ↓	Acc ↑	W ↓	Acc ↑	W ↓	Win Rate ↑
Lower bound for W	2.0	—	1.5	—	1.1	—	0.26	—	0.25	—	0.26	—
Qwen3-4B	26.9	21.7	13.1	27.9	20.1	24.3%	1.14	37.0%	0.65	37.2%	0.66	9%
Llama3.3-70B	24.2	27.0	15.3	26.4	14.9	28.4%	1.02	37.7%	0.59	43.9%	0.47	6%
HY3	30.0	23.0	15.8	27.1	14.0	27.0%	0.80	42.1%	0.46	45.6%	0.41	10%
DeepSeek V3.2	16.9	20.9	10.1	27.5	12.3	27.2%	0.76	36.2%	0.54	40.6%	0.47	14%
GLM-5.2	20.4	18.8	10.1	25.7	18.7	27.0%	0.77	41.2%	0.45	45.9%	0.38	16%
Claude Sonnet 4.6	24.6	18.7	10.6	25.3	19.3	28.6%	0.86	41.8%	0.45	46.7%	0.36	**17%**
Claude Opus 4.6	26.0	19.3	11.7	25.9	20.1	29.7%	0.86	41.8%	**0.43**	45.1%	0.37	15%
Claude Opus 5	26.7	**17.7**	11.5	25.8	15.4	30.7%	0.73	40.6%	0.48	**49.0%**	**0.34**	16%
GPT 4.1	27.5	18.1	**6.9**	27.1	16.3	29.2%	0.79	39.0%	0.48	46.4%	0.38	**17%**
GPT 5.4	31.4	21.1	15.2	26.1	14.0	29.6%	0.90	42.4%	0.46	47.3%	0.38	8%
GPT 5.6 Terra	25.0	21.7	14.0	27.2	14.9	30.7%	0.90	41.0%	0.48	47.3%	0.40	11%
GPT 5.6 Sol	24.9	22.0	16.1	25.5	13.4	**32.4%**	0.88	**43.1%**	0.46	48.0%	0.40	10%
Gemini 3.1 Flash	31.3	25.5	18.6	29.4	17.3	30.5%	0.82	40.5%	0.52	46.3%	0.41	14%
Gemini 3.1 Pro	21.3	19.4	12.5	26.7	**10.3**	30.5%	0.75	42.5%	0.46	48.5%	0.36	12%
Socrates-Llama3-8B-SFT	**16.3**	39.5	21.3	37.7	14.6	24.0%	0.55	23.5%	0.62	24.0%	0.67	2%
Socrates-Llama3-8B-DPO	26.4	27.3	19.7	27.7	21.4	24.7%	0.93	25.9%	0.96	28.4%	0.76	3%
Socrates-Qwen-14B-SFT	17.9	35.9	15.9	36.7	14.7	23.8%	0.69	23.8%	0.71	26.2%	0.69	2%
Socrates-Qwen-14B-DPO	20.0	20.5	14.4	**24.8**	17.7	23.9%	1.01	35.3%	0.71	37.8%	0.56	8%
Centaur-70B	16.9	23.3	11.1	29.3	13.9	24.7%	**0.43**	33.1%	0.48	36.1%	0.49	4%
OSim-8B	21.9	22.7	12.4	29.4	21.2	29.4%	0.90	37.5%	0.63	37.8%	0.60	14%

Capability	Subject-trait Inference						Knowledge Application	
Task	Demo. To Pers.		Across-Dim Pers. Pred.		Pers. To Demo.		Sci. Workflow Pred.	Econ. Contest
Metric	MAE↓	W↓	MAE↓	W↓	MAE↓	W↓	BLEURT↑	Acc. ↑
Lower bound for W	—	0.65	—	0.65	—	0.42	—	—
Qwen3-4B	7.3	5.14	7.6	5.21	11.2	10.44	0.450	73.9% ± 1.3
Llama3.3-70B	6.9	5.18	8.3	5.19	9.3	7.67	0.430	65.8% ± 0.6
HY3	7.1	4.91	7.7	5.32	13.9	7.21	0.474	91.3% ± 0.9
DeepSeek V3.2	7.3	5.12	7.5	4.61	10.7	7.57	0.430	74.4% ± 1.6
GLM-5.2	7.0	5.47	7.4	4.51	10.5	4.27	0.458	92.7% ± 1.4
Claude Sonnet 4.6	7.6	5.83	7.2	4.37	14.5	8.72	0.472	83.1% ± 1.3
Claude Opus 4.6	7.0	5.18	6.7	4.49	13.3	8.47	**0.477**	**95.6%** ± 0.6
Claude Opus 5	**6.3**	4.96	**6.0**	3.65	**9.1**	3.70	0.440	95.2% ± 0.0
GPT 4.1	6.9	4.92	7.2	3.93	10.6	4.01	0.460	78.4% ± 0.8
GPT 5.4	7.4	5.70	7.3	4.40	10.5	4.05	0.462	**95.6%** ± 0.6
GPT 5.6 Terra	7.5	6.73	7.2	4.59	10.0	3.14	0.465	92.7% ± 1.0
GPT 5.6 Sol	7.6	6.84	7.0	4.39	11.5	7.19	0.468	94.8% ± 0.6
Gemini 3.1 Flash	7.1	4.78	7.1	4.98	14.0	8.17	0.432	77.6% ± 1.7
Gemini 3.1 Pro	6.4	4.38	6.8	**3.60**	10.5	**1.67**	0.470	**95.6%** ± 0.4
Socrates-Llama3-8B-SFT	13.3	10.38	12.4	9.47	12.6	4.81	0.026	0.0% ± 0.0
Socrates-Llama3-8B-DPO	8.7	6.14	8.6	6.73	9.6	5.97	0.426	39.7% ± 4.6
Socrates-Qwen-14B-SFT	9.0	3.77	9.8	4.90	11.0	3.69	0.024	0.0% ± 0.0
Socrates-Qwen-14B-DPO	8.1	6.46	9.0	6.85	10.5	7.83	0.448	70.3% ± 0.6
Centaur-70B	9.8	**3.23**	9.3	3.83	10.6	5.18	0.430	37.7% ± 2.8
OSim-8B	7.0	5.69	7.5	5.13	10.9	9.04	0.449	68.5% ± 1.4

Table 3: Benchmarking results on `BehaviorBench`. **Bold** and underline denote the best and second best performance.

Behavioral foundation models. This category includes foundation models fine-tuned on behavior-related data. In the `BehaviorBench` evaluation, we include three representative model families. Centaur-70B (Binz et al., 2024) is fine-tuned on a cognitive-psychology corpus, while Socrates (Kolluri et al., 2025) is fine-tuned on social-science survey data. OSim-8B (Zhou et al., 2026) is fine-tuned on diverse human interaction data, including conversation, social simulation, and role-play.

3.2 BehaviorBench Leaderboards

Beyond heterogeneous per-task metrics, we aim to derive a unified score to compare models across tasks. We treat each task as a pairwise comparison among the models, inspired by the Holistic Evaluation of Language Models (HELM) leaderboards (Liang et al., 2022). Concretely, we define a model's *win rate* as the fraction of other models it outperforms on a given task (based on the task-specific metric), and *mean win rate* as the average of win rates across tasks. The leaderboards include the mean win rate rankings at both the individual and distributional levels.

4 RESULTS

4.1 MODEL CAPABILITIES

Table 3 reports model performance across `BehaviorBench` tasks, grouped by capabilities. We summarize the main findings below. Appendix F reports per-game results for the game tasks.

`BehaviorBench` remains challenging for frontier models. At the individual level, the best accuracies on Demo. To Resp., Seq. Resp. Pred., and Masked Resp. Pred. are only 32.4%, 43.1%, and 49.0% respectively. On Multi-Round Pred. and Across-Ctx Pred. tasks, even the best performance is off by 17.7 and 24.8 in MAE, when the range is 0–100. For strategic game play, the best models only win 17% of the Beauty Contest sessions they play. On distributional metrics, we follow previous work Suh et al. (2025); Xie et al. (2026) and compute a lower bound on W by bootstrapping the human responses (first row of Table 3). We find that the best W on each task is 4.6 times larger than this lower bound on average, indicating there is still substantial room for improvement.

	$\sigma_{\text{model}}/\sigma_{\text{human}}$	$\frac{\lvert\mu_{\text{model}}-\mu_{\text{human}}\rvert}{\sigma_{\text{human}}}\downarrow$
Gemini 3.1 Pro	<u>0.70</u>	<u>0.31</u>
Centaur-70B	**0.93**	0.38
Claude Opus 5	0.53	**0.26**
GPT 4.1	0.62	0.38
GLM-5.2	0.61	0.34
Claude Opus 4.6	0.54	0.42
GPT 5.6 Sol	0.63	0.46

Table 4: Standard-deviation ratio and standardized bias of model predictions, averaged across the nine tasks with distributional evaluation metrics. Best in each column is **bold** (deviation ratio closest to 1 and lowest bias), second best <u>underlined</u>.

Gemini 3.1 Pro and Claude Opus 5 lead `BehaviorBench`. As shown in Figure 1, Claude Opus 5 ranks 1st of the 20 models at the individual level and 3rd at the distributional level, and Gemini 3.1 Pro ranks 2nd and 1st. They are the only two models in the top three at both levels.

Frontier proprietary models retain an advantage on contest-style economic reasoning. Claude Opus 4.6, GPT 5.4, and Gemini 3.1 Pro each achieve 95.6% accuracy on the economics contest task. In contrast, Centaur-70B and most Socrates variants perform notably worse on knowledge-intensive tasks, especially the economics contest, suggesting that models fine-tuned for behavioral prediction do not automatically acquire broader behavioral-science reasoning ability.

4.2 INDIVIDUAL-LEVEL AND DISTRIBUTIONAL PERFORMANCE DO NOT ALWAYS ALIGN

Model performance at the individual level does not necessarily guarantee a strong distributional performance. For example, Claude Opus 4.6 and GPT 5.6 Sol rank 3rd and 4th at the individual level, but fall to 9th and 11th at the distributional level. By contrast, Centaur-70B ranks 2nd at the distributional level, but only ranks 18th at the individual level. To examine how behavioral foundation models better preserve the diversity of human responses than general-purpose models, we report two statistics for models that rank among the top five on either the individual-level or distributional leaderboard:

- $\sigma_{\text{model}}/\sigma_{\text{human}}$, which compares the variation in model predictions with that in human responses. A value closer to 1 indicates that the model better captures the diversity of human responses.
- $\frac{\lvert\mu_{\text{model}}-\mu_{\text{human}}\rvert}{\sigma_{\text{human}}}$, which measures the difference between the model and human population means, normalized by the standard deviation of human responses. A lower value indicates better alignment with the human population mean.

We average both statistics across the nine distributional metrics in `BehaviorBench`. As shown in Table 4, **general-purpose LLMs tend to underestimate the diversity of human responses, even though the leading ones predict the population mean more accurately. Behavioral foundation models, by contrast, better preserve response diversity**. The general-purpose models in Table 4 reproduce only 0.53–0.70 of the human standard deviation, against 0.93 for Centaur-70B.

5 CO-OPTIMIZE INDIVIDUAL-LEVEL AND DISTRIBUTIONAL PERFORMANCE

In Section 4.2, we find that general-purpose LLMs and behavior foundation models exhibit uneven performance at distributional and individual levels. In this section, we ask the question: is it possible to optimize performance on these two evaluation levels at the same time? To answer this question, we fine-tune general-purpose LLMs on diverse behavior data targeting the capabilities in `BehaviorBench`.

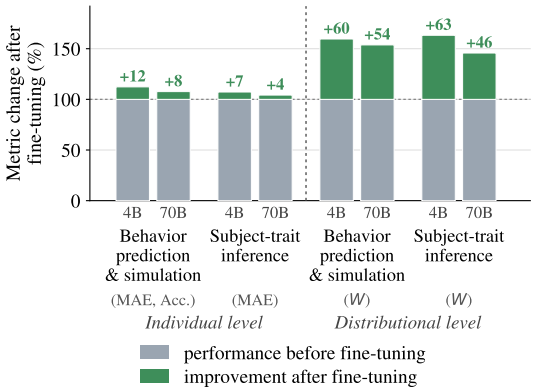


Figure 2: Performance improvement after fine-tuning. Both Qwen3-4B and Llama-3.3-70B improve more at the distributional level than the individual level.

5.1 TRAINING DATA AND SETUP

We fine-tune two widely used open-weight backbone models, Qwen3-4B (Team, 2025) and Llama-3.3-70B (Team, 2024), on diverse behavioral data. Following the task formulation in Section 2.2, we instantiate fine-tuning tasks that target the capabilities evaluated by `BehaviorBench`.

Compared with the fine-tuning datasets of previous works on behavior foundation models, our curated dataset prioritizes breadth of data sources over depth within each source. Specifically, we include all three data sources described in Section 2.1. In contrast, Centaur-70B (Binz et al., 2024) uses only experimental data, while Socrates (Kolluri et al., 2025) uses only survey data. In particular, our fine-tuning dataset includes 649,394 data points across 11 fine-tuning tasks. We note that the training data is disjoint from `BehaviorBench`. Details on the backbone models, the fine-tuning schema, and the fine-tuning data are described in Appendix D.

5.2 IMPROVEMENT ON DISTRIBUTIONAL LEVEL IS LARGER THAN INDIVIDUAL LEVEL

As shown in Figure 2, both fine-tuned models improve over their respective backbone models. The gains are bigger at the distributional level (46%–63%), while individual-level prediction improves consistently but by much smaller margins (4.0%–12.3%). One explanation is that the two levels involve different prediction targets. Distributional metrics compare only the predicted behavior distribution with the human distribution, $p(y \mid c)$, whereas individual-level metrics require accurate subject-specific predictions, $p(y \mid x, c)$. Our experiment shows that fine-tuning on human behavior data may improve predictions of the population distribution more than predictions for individual subjects. Overall, the strong performance of fine-tuned models implies that individual-level accuracy and faithful distributional alignment can be improved simultaneously, while also revealing substantial room for improvement.

6 CONCLUSION

We introduce `BehaviorBench`, a comprehensive benchmark that evaluates foundation models across four capabilities with metrics at both individual and distributional levels. Our evaluation reveals that current foundation models exhibit uneven strengths across tasks. Particularly, general-purpose frontier proprietary LLMs excel at individual-level prediction and knowledge-intensive tasks. In contrast, behavioral foundation models perform strongly on distributional alignment. Our fine-tuning experiment suggests that models can achieve both individual-level accuracy and faithful distributional alignment, while showing that foundation models still have substantial room to improve for behavioral science. More broadly, by emphasizing distributional alignment, `BehaviorBench` can help assess whether AI systems reflect the heterogeneity of human populations rather than overfitting to majority or average behaviors. Our results establish `BehaviorBench` as a foundation for developing and assessing behaviorally aligned AI systems.

Ethics Statement

This work uses previously collected human behavioral and survey data and does not involve new participant recruitment or data collection. We use these data in accordance with the licenses and terms described in Appendix I, and we do not attempt to identify individual participants. Because the datasets contain behavioral histories and demographic attributes, they may raise privacy concerns and may not represent all populations, cultures, or behavioral contexts.

Models that simulate human behavior may reproduce stereotypes, underestimate population diversity, or produce inaccurate predictions about particular groups or individuals. `BehaviorBench` is intended for research evaluation and should not be used to replace human-subject studies or to support consequential decisions about individuals.

AI Use Statement

In this work, we used generative AI tools to support data analysis and to reformat the literature data. We have not used generative AI tools for any other task that requires disclosure. Additionally, we used generative AI tools to aid and polish the writing of this paper and to create and edit software code. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility Statement

The supplementary material contains the anonymized evaluation code of `BehaviorBench` and the evaluation data of every task. Section 2.1 and Appendix B describe the data sources, cleaning, and splits. Table 1 lists each task with its evaluation-set size and metrics. Appendices C, D, and H give the metric details, the fine-tuning settings, and the prompt templates.

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 Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pp. 337–371. PMLR, 2023. URL https://proceedings.mlr.press/v202/aher23a.html.

Anthropic. Introducing claude opus 4.6. Anthropic Blog, 2026a. URL https://www.anthropic.com/news/claude-opus-4-6. February 5, 2026.

Anthropic. Introducing claude opus 5. Anthropic Blog, 2026b. URL https://www.anthropic.com/news/claude-opus-5. July 24, 2026.

Anthropic. Introducing claude sonnet 4.6. Anthropic Blog, 2026c. URL https://www.anthropic.com/news/claude-sonnet-4-6. February 17, 2026.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Christopher A Bail. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121, 2024.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. Centaur: a foundation model of human cognition. *arXiv preprint arXiv:2410.20268*, 2024.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32 (4):401–416, 2024.

Wiebke Bleidorn, Patrick L Hill, Mitja D Back, Jaap JA Denissen, Marie Hennecke, Christopher J Hopwood, Markus Jokela, Christian Kandler, Richard E Lucas, Maike Luhmann, et al. The policy relevance of personality traits. *American psychologist*, 74(9):1056, 2019.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. The emergence of economic rationality of gpt. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120, 2023.

DeepSeek-AI. Deepseek-v3.2: Pushing the frontier of open large language models. *CoRR*, abs/2512.02556, 2025. doi: 10.48550/ARXIV.2512.02556. URL https://doi.org/10.48550/arXiv.2512.02556.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701, 2023.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120, 2023.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations. *CoRR*, abs/2402.12348, 2024. doi: 10.48550/ARXIV.2402.12348. URL https://doi.org/10.48550/arXiv.2402.12348.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.

Herbert Gintis. *The bounds of reason: game theory and the unification of the behavioral sciences-revised edition*. Princeton university press, 2014.

GLM. GLM-5: from vibe coding to agentic engineering. *CoRR*, abs/2602.15763, 2026. doi: 10.48550/ARXIV.2602.15763. URL https://doi.org/10.48550/arXiv.2602.15763.

Google. Gemini 3.1 pro: A smarter model for your most complex tasks. Google Blog, 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/. February 19, 2026.

Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. Ai and the transformation of social science research. *Science*, 380(6650): 1108–1109, 2023.

Shangmin Guo, Haoran Bu, Haochuan Wang, Yi Ren, Dianbo Sui, Yuming Shang, and Siting Lu. Economics arena for large language models. *CoRR*, abs/2401.01735, 2024. doi: 10.48550/ARXIV.2401.01735. URL https://doi.org/10.48550/arXiv.2401.01735.

Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. Predicting results of social science experiments using large language models. *Preprint*, 2024.

Teck-Hua Ho, Colin Camerer, and Keith Weigelt. Iterated dominance and iterated best response in experimental" p-beauty contests". *The American Economic Review*, 88(4):947–969, 1998.

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

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. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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. *CoRR*, abs/2510.17516, 2025. doi: 10.48550/ARXIV.2510.17516. URL https://doi.org/10.48550/arXiv.2510.17516.

Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael R. Lyu. Competing large language models in multi-agent gaming environments. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=DI4gW8viB6.

Matthew O. Jackson, Qiaozhu Mei, Stephanie W. Wang, Yutong Xie, Walter Yuan, Seth Benzell, Erik Brynjolfsson, Colin F. Camerer, James Evans, Brian Jabarian, Jon M. Kleinberg, Juanjuan Meng, Sendhil Mullainathan, Asuman Ozdaglar, Thomas Pfeiffer, Moshe Tennenholtz, Robb Willer, Diyi Yang, and Teng Ye. AI behavioral science. *CoRR*, abs/2509.13323, 2025. doi: 10.48550/ARXIV.2509.13323. URL https://doi.org/10.48550/arXiv.2509.13323.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S. Bernstein. Finetuning llms for human behavior prediction in social science experiments. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 30096–30111. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1530. URL https://doi.org/10.18653/v1/2025.emnlp-main.1530.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *CoRR*, abs/2211.09110, 2022. doi: 10.48550/ARXIV.2211.09110. URL https://doi.org/10.48550/arXiv.2211.09110.

Steffen Lippert, Anna Dreber, Magnus Johannesson, Warren Tierney, Wilson Cyrus-Lai, Eric Luis Uhlmann, Thomas Pfeiffer, Emotion Expression Collaboration, et al. Can large language models help predict results from a complex behavioural science study? *Royal Society Open Science*, 11(9), 2024.

Hongtao Liu, Zhicheng Du, Zihe Wang, and Weiran Shen. Chbench: A cognitive hierarchy benchmark for evaluating strategic reasoning capability of llms. *CoRR*, abs/2508.11944, 2025. doi: 10.48550/ARXIV.2508.11944. URL https://doi.org/10.48550/arXiv.2508.11944.

Nunzio Lorè and Babak Heydari. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14(1):18490, 2024.

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.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. Benchmarking distributional alignment of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 24–49. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.2. URL https://doi.org/10.18653/v1/2025.naacl-long.2.

Juanjuan Meng. Ai emerges as the frontier in behavioral science. *Proceedings of the National Academy of Sciences*, 121(10):e2401336121, 2024.

Rosemarie Nagel. Unraveling in guessing games: An experimental study. *The American economic review*, 85(5):1313–1326, 1995.

OpenAI. Introducing gpt-4.1 in the api. OpenAI Blog, 2025a. URL https://openai.com/index/gpt-4-1/. April 14, 2025.

OpenAI. Introducing gpt-5. OpenAI Blog, 2025b. URL https://openai.com/index/introducing-gpt-5/. August 7, 2025.

OpenAI. Introducing gpt-5.4. OpenAI Blog, 2026a. URL https://openai.com/index/introducing-gpt-5-4/. March 5, 2026.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. OpenAI Blog, 2026b. URL https://openai.com/index/gpt-5-6/. July 9, 2026.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In Sean Follmer, Jeff Han, Jürgen Steimle, and Nathalie Henry Riche (eds.), *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pp. 2:1–2:22. ACM, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie J. Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative agent simulations of 1,000 people. *CoRR*, abs/2411.10109, 2024. doi: 10.48550/ARXIV.2411.10109. URL https://doi.org/10.48550/arXiv.2411.10109.

Barbara Rychalska, Szymon Lukasik, and Jacek Dabrowski. Synerise monad: A foundation model for behavioral event data. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (eds.), *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pp. 3344–3348. ACM, 2023. doi: 10.1145/3539618.3591851. URL https://doi.org/10.1145/3539618.3591851.

Aadesh Salecha, Molly E. Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H. Ungar, and Johannes C. Eichstaedt. Large language models show human-like social desirability biases in survey responses. *CoRR*, abs/2405.06058, 2024. doi: 10.48550/ARXIV.2405.06058. URL https://doi.org/10.48550/arXiv.2405.06058.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pp. 29971–30004. PMLR, 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. BLEURT: learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 7881–7892. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.704. URL https://doi.org/10.18653/v1/2020.acl-main.704.

Eilam Shapira, Omer Madmon, Itamar Reinman, Samuel Joseph Amouyal, Roi Reichart, and Moshe Tennenholtz. GLEE: A unified framework and benchmark for language-based economic environments. *CoRR*, abs/2410.05254, 2024. doi: 10.48550/ARXIV.2410.05254. URL https://doi.org/10.48550/arXiv.2410.05254.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, 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 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 21147–21170. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.1028/.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Tencent Hy Team. Hy3, 2026. URL https://huggingface.co/tencent/Hy3.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Jia Wang, Ziyu Zhao, Tingjuntao Ni, and Zhongyu Wei. Sociobench: Modeling human behavior in sociological surveys with large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 26257–26289. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1335. URL https://doi.org/10.18653/v1/2025.emnlp-main.1335.

Patrick Y Wu, Jonathan Nagler, Joshua A Tucker, and Solomon Messing. Large language models can be used to scale the ideologies of politicians in a zero-shot learning setting. *preprint*, 2023.

Yutong Xie, Yiyao Liu, Zhuang Ma, Lin Shi, Xiyuan Wang, Walter Yuan, Matthew O. Jackson, and Qiaozhu Mei. How different AI chatbots behave? benchmarking large language models in behavioral economics games. *CoRR*, abs/2412.12362, 2024. doi: 10.48550/ARXIV.2412.12362. URL https://doi.org/10.48550/arXiv.2412.12362.

Yutong Xie, Zhuoheng Li, Xiyuan Wang, Yijun Pan, Qijia Liu, Xingzhi Cui, Kuang-Yu Lo, Ruoyi Gao, Xingjian Zhang, Jin Huang, Walter Yuan, Matthew O. Jackson, and Qiaozhu Mei. Be.fm: Open foundation models for human behavior. *CoRR*, abs/2505.23058, 2025. doi: 10.48550/ARXIV.2505.23058. URL https://doi.org/10.48550/arXiv.2505.23058.

Yutong Xie, Ruoyi Gao, and Qiaozhu Mei. Distributional alignment for social simulation with llms: A mixture modeling approach. In Won Kim, Jae-Gil Lee, Kyuseok Shim, James Caverlee, David F. Gleich, Xiangliang Zhang, Qi He, Faisal Farooq, Rómer Rosales, Sayan Ranu, Haixun Wang, Yan Liu, Carl Yang, Leman Akoglu, Nitesh V. Chawla, and Vagelis Papalexakis (eds.), *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD 2026, Jeju Island, Korea, August 9-13, 2026*, pp. 12492–12503. ACM, 2026. doi: 10.1145/3770855.3818919. URL https://doi.org/10.1145/3770855.3818919.

Kai-Cheng Yang and Filippo Menczer. Large language models can rate news outlet credibility. *CoRR*, abs/2304.00228, 2023. doi: 10.48550/ARXIV.2304.00228. URL https://doi.org/10.48550/arXiv.2304.00228.

Yanwu Yang and Panyu Zhai. Click-through rate prediction in online advertising: A literature review. *Information Processing & Management*, 59(2):102853, 2022.

Xingjian Zhang, Yutong Xie, Jin Huang, Jinge Ma, Zhaoying Pan, Qijia Liu, Ziyang Xiong, Tolga Ergen, Dongsub Shim, Honglak Lee, and Qiaozhu Mei. MASSW: A new dataset and benchmark tasks for ai-assisted scientific workflows. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, Findings of ACL, pp. 2373–2394. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-NAACL.127. URL https://doi.org/10.18653/v1/2025.findings-naacl.127.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT: A scalable lightweight infrastructure for fine-tuning. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 29733–29735. AAAI Press, 2025. doi: 10.1609/AAAI.V39I28.35383. URL https://doi.org/10.1609/aaai.v39i28.35383.

Xuhui Zhou, Weiwei Sun, Weihua Du, Jiarui Liu, Haojia Sun, Qianou Ma, Tongshuang Wu, Yiming Yang, and Maarten Sap. OdysSim: Building foundation models for human behavior simulation. *CoRR*, abs/2606.14199, 2026. doi: 10.48550/ARXIV.2606.14199. URL https://doi.org/10.48550/arXiv.2606.14199.

A Related Work

A.1 Behavioral Models and Foundation Models for Behavioral Science

General foundation models, including GPT and Llama, are increasingly being applied to behavioral science domains such as psychology, sociology, and economics (Demszky et al., 2023; Meng, 2024; Bail, 2024). These models have shown promise in predicting survey responses (Argyle et al., 2023; Grossmann et al., 2023), enabling automated content analysis (Wu et al., 2023; Yang & Menczer, 2023; Gilardi et al., 2023), supporting agent-based modeling (Park et al., 2023; 2024), and inferring experimental results (Hewitt et al., 2024; Lippert et al., 2024). However, recent work has also revealed important limitations in directly applying general-purpose LLMs to behavioral science tasks. For example, these models often fail to capture population variance (Bisbee et al., 2024), exhibit sensitivity or bias toward specific data formats (Dentella et al., 2023), and underrepresent certain populations (Santurkar et al., 2023). This mismatch between the needs of behavioral science and the design of general foundation models calls for domain-specific models that better align with behavioral science principles, data, and evaluation goals.

A small but growing body of work has begun to address this need. Monad (Rychalska et al., 2023), for example, learns dense user representations from streams of clicks, transactions, and other behavioral events. Centaur (Binz et al., 2024) models human responses in cognitive-psychology experiments, including tasks involving decision-making, memory, supervised learning, and Markov decision processes. `Be.FM` models (Xie et al., 2025), trained on experimental, survey, and literature data, demonstrate capabilities in predicting human behavior, inferring subject characteristics, generating insights about contextual factors, and applying behavioral science knowledge to problem-solving. Socrates (Kolluri et al., 2025) adapts language models to social-science survey data. However, despite this growing body of work, we still lack a systematic understanding of how well foundation models perform across diverse behavioral science tasks, contexts, and populations.

A.2 AI Behavioral Science

AI behavioral science is inherently bidirectional—not only can AI serve as a tool in behavioral science research as discussed above, but behavioral science methods can also be used to study the behaviors, biases, and emergent properties of AI systems themselves (Jackson et al., 2025). For example, Aher et al. (2023); Mei et al. (2024) and Xie et al. (2024) have conducted variations on Turing tests, comparing LLMs' decisions in economic games to human data. Similarly, Horton et al. (2023); Lorè & Heydari (2024) investigated the strategic behavior of LLMs, while Chen et al. (2023) assessed GPT's economic rationality. Other studies applied behavioral frameworks such as theory of mind to LLMs (Strachan et al., 2024), or examined biases in AI-generated survey responses (Salecha et al., 2024).

While these two directions—using AI for behavioral science and applying behavioral science to AI—have each advanced, a unifying approach remains missing. Specifically, there is a lack of models that function both as tools for understanding human behavior and as platforms for behavioral interrogation of AI systems themselves.

A.3 Benchmarks for Behavior Modeling and Economic Reasoning

A growing line of work has developed benchmarks for evaluating how well language models can reproduce human responses in surveys and social science experiments. OpinionQA (Santurkar et al., 2023) compares model outputs against U.S. public opinion surveys to study whose opinions language models reflect across demographic groups. SimBench (Hu et al., 2025) harmonizes 20 datasets spanning moral dilemmas, economic choices, and psychological assessments into a single standardized benchmark, and measures how closely model predictions match group-level human response distributions. SubPOP (Suh et al., 2025) targets the prediction of public opinion distributions, showing that fine-tuning on survey data substantially improves alignment with held-out human responses. SocioBench (Wang et al., 2025) models human behavior in sociological surveys, while SocSci210 (Kolluri et al., 2025) fine-tunes language models on social science experiments to predict participant-level and distributional-level outcomes. These benchmarks have established survey response prediction as a central task for evaluating LLMs as proxies for human respondents. However,

most prior work focuses primarily on this single task formulation without exploring alternative formulations such as predicting missing survey responses.

Beyond predicting responses in surveys and social science experiments, recent works also evaluate LLMs in economic and strategic settings. GTBench (Duan et al., 2024) evaluates the strategic reasoning limitations of LLMs across a suite of classic game-theoretic scenarios, and GAMA-Bench (Huang et al., 2025) embeds competing language models within multi-agent gaming environments. CHBench (Liu et al., 2025) grounds evaluation in cognitive hierarchy theory, while Economics Arena (Guo et al., 2024) benchmarks LLMs across a range of economic decision tasks. GLEE (Shapira et al., 2024) provides a unified framework and benchmark for language-based economic environments such as bargaining, negotiation, and persuasion games. Most of these benchmarks have LLMs play against other LLMs or against heuristics, with no human opponents. In our work, we use human play records from economic games to construct a proxy evaluation for LLMs playing against human players.

B Data Details

B.1 Data Sources

This subsection provides a more detailed walkthrough of the data sources introduced in Section 2.1.

Experimental data. Experimental data document the behavioral patterns of varied participant groups under controlled conditions, providing crucial insights into how individuals and groups make decisions and engage socially. We utilize a collection of experiments from MobLab, a platform for behavioral science and economic games. After cleaning, this source contains 379,754 behavioral observations from 76,035 unique subjects collected over a nine-year period (2015 to 2023) (Mei et al., 2024). The collection covers seven classic economic games across nine scenarios: Dictator, Ultimatum (Proposer and Responder), Trust (Investor and Banker), Public Goods, Bomb Risk, Beauty Contest, and Push/Pull (Prisoner's Dilemma).

Survey data. Survey data offer self-reported information about personality traits, attitudes, and preferences. They primarily characterize subject attributes (x in Equation 1), though they may also contain self-reported behaviors $\tilde{y}$, which serve as a proxy for observed behaviors y and can be leveraged to simulate responses to surveys not yet conducted. Such responses may be elicited either without any specified context or under particular pre-defined scenarios c embedded in the survey items. We adopt the publicly available Big Five Personality Test dataset[4], which contains demographic attributes (race, age, and gender) of participants along with their single-choice responses to 50 items spanning five personality dimensions: openness, conscientiousness, extraversion, agreeableness, and neuroticism. The dataset comprises 19,630 subjects and a total of 981,500 survey responses.

Literature data. The behavioral science literature captures established theories and empirical findings on human behavior, serving as a proxy for the behavioral knowledge $\mathbb{K}$ in Equation 1. We gather titles and abstracts from two leading interdisciplinary venues for behavioral research: the *American Economic Review* (AER) and *Nature Human Behaviour* (NHB). The resulting corpus consists of 6,515 AER articles spanning 1911 to 2025 and 871 NHB articles spanning 2016 to 2025. Following the approach of MASSW (Zhang et al., 2025), we summarize each title–abstract pair into a structured research workflow capturing the research context, key idea, method, outcome, and anticipated impact using GPT-5 (OpenAI, 2025b).

Economics contest data. We also use multiple-choice questions from the International Economics Olympiad (IEO) as an evaluation-only source for economics problem solving. This source differs from the experimental and survey datasets because it does not record human behavioral variation. Instead, it evaluates whether models can represent, reason about, and apply domain-specific economic knowledge to unseen problems.

[4] https://www.kaggle.com/datasets/lucasgreenwell/ocean-five-factor-personality-test-responses.

B.2 Data Quality Control

Each data source passes a cleaning and validation pipeline before any task instance is constructed. We report the filters and the retention rate for each source below.

MobLab economic games. Raw move records pass a four-level filter: (L1) records carrying an invalid placeholder value in the move or round result, together with all subsequent rounds of the same subject; (L2) moves outside the action range implied by the game instruction; (L3) sessions with a non-standard game configuration (endowment, group size, or role), excluded so that the payoff structure is comparable across subjects; and (L4) rounds that create a gap in a subject's round sequence after L1–L3. Of the 569,795 raw records, 379,754 (66.6%) remain, covering 76,035 of 85,343 subjects.

Big Five survey. We remove one respondent who left all 50 items blank and 88 respondents reporting an implausible age (≥ 80), retaining 19,630 of 19,719 respondents (99.5%). The retained respondents span 158 countries with a median age of 22.

Literature corpus. Of 9,255 AER records retrieved from OpenAlex, 6,515 have both a title and an abstract and are retained; all 871 NHB records have both. The GPT-5 workflow extraction returns a set of five aspects for $6{,}514/6{,}515$ AER and $870/871$ NHB papers. After this filter, the 257 papers published in 2025 yield 1,084 of 1,285 possible evaluation instances.

Economics contest. The source contains 130 multiple-choice problems across seven IEO years (2018, 2019, and 2021–2025). We remove the 6 problems whose answer depends on a figure that cannot be reproduced in text, leaving the 124 problems used for evaluation.

B.3 Data Splits

This subsection describes how each raw source from Section B.1 is partitioned into the fine-tuning set and the `BehaviorBench` evaluation set, and how we sample task instances from each partition. Table 5 summarizes the per-task entry counts and the split rule used. Across all sources, the partition is constructed so that no subject (or paper) appears on both the fine-tuning and evaluation side of any task.

MobLab economic games. We split unique participants (identified by a unique `UserID`) 70/10/20 into train, validation, and test, so a participant's complete history of game play remains within a single partition. The Game Behavior Sim., Multi-Round Pred., Across-Ctx Pred., and Strategic Game Play tasks are all generated from these splits. Raw game logs are first cleaned by a filtering pipeline that removes invalid game choices.

The Push/Pull game (held-out). The Push/Pull game is excluded from fine-tuning entirely. When constructing the Push/Pull evaluation, we additionally drop any subject that appears in the MobLab train or validation split, so no participant the model has seen during fine-tuning contributes to this held-out evaluation.

Big Five survey. We use a 90/10 split over subjects. All Big Five tasks (Demo. to Resp., Seq. Resp. Pred., Masked Resp. Pred., Demo. to Pers., Across-Dim Pers. Pred., and Pers. to Demo.) draw their fine-tuning and evaluation instances from this single participant-level split.

Scientific workflows. We split AER and NHB papers by publication year. Papers published in 2025 form the workflow-prediction test set, and all earlier years form the fine-tuning set. This temporal cut reflects the practical use case of generalizing to newly published papers.

Economics contest (IEO). The IEO multiple-choice problems serve as an evaluation-only source and have no corresponding fine-tuning data.

Capability	Task	# Train	# Eval	Train/Test Split
Behavior prediction and simulation	Game Behav. Sim.	78.8K	1,800	by subject
	Multi-Round Pred.	57.9K	3,623	by subject
	Across-Ctx Pred.	44.1K	6,262	by subject
	Demo. To Resp.	88.3K	1,000	by subject
	Seq. Resp. Pred.	36.0K	1,000	by subject
	Masked Resp. Pred.	40.0K	1,000	by subject
Subject-trait inference	Demo. To Pers.	88.3K	1,000	by subject
	Across-Dim Pers. Pred.	88.3K	1,000	by subject
	Pers. To Demo.	17.7K	1,000	by subject
Strategic decision-making	Strategic Game Play	85.2K	1,000	by subject
Behavioral knowledge application	Sci. Workflow Pred.	24.7K	1,084	by publication year
	Economics Contest	—	124	eval only
Total		**649.4K**	**19,893**	

Table 5: Training and evaluation data summary.

Sampling and downsampling. For tasks where the natural enumeration produces far more instances than we need for either fine-tuning or evaluation, we downsample to a fixed target size. All Big Five evaluation tasks are capped at 1K examples. MobLab game tasks use random sampling per game, with a cap of 200, 500, and 750 examples per game for Game Behav. Sim., Multi-Round Pred., and Across-Ctx Pred. respectively. The Strategic Game Play evaluation set is capped at 1K examples. The final per-task counts after sampling are reported in Table 5.

B.4 Evaluation Tasks in BehaviorBench

For each evaluation task in `BehaviorBench`, we briefly describe the task below and provide its prompt template.

Single-round behavior simulation (Game Behav. Sim.). The model is shown the rules of a single MobLab economic game and asked for its first-round move. The game-specific instructions cover the Dictator game, the Ultimatum game (as Proposer or Responder), the Trust game (as Investor or Banker), the Public Goods game, the Bomb Risk game, the Beauty Contest game, and the Push/Pull game (Prisoner's Dilemma) ([PROMPT]).

Multi-round behavior prediction (Multi-Round Pred.). The model is shown the game instruction together with the subject's prior rounds in the same session, and predicts the next round's move ([PROMPT]).

Across-context single-round prediction (Across-Ctx Pred.). The model is shown a subject's first-round moves in several other games and asked to predict the subject's first-round move in a target game ([PROMPT]).

Survey response prediction given demographics (Demo. to Resp.). The model is given a subject's demographics and asked to predict one Big Five item of this subject on a five-point Likert scale ([PROMPT]).

Sequential survey response prediction (Seq. Resp. Pred.). The model is shown the subject's previous answers in the original survey order (interleaved across the five dimensions) and predicts the subject's response to the next item ([PROMPT]).

Masked survey response prediction (Masked Resp. Pred.). The model is shown 49 of a subject's 50 Big Five item responses, grouped by dimension, and predicts the held-out item ([PROMPT]).

Personality score from demographics (Demo. to Pers.). The model is given a subject's demographics and predicts the subject's score on one Big Five dimension ([PROMPT]).

Across-dimension personality prediction (Across-Dim Pers. Pred.). The model is given a subject's scores on four Big Five dimensions and predicts the score on the held-out dimension ([PROMPT]).

Age prediction from personality (Pers. to Demo.). The model is given a subject's scores on all five Big Five dimensions and predicts the subject's age ([PROMPT]).

Strategic game play. The model plays the Beauty Contest game against human players. Given the play history and the group average from each prior round, the model picks an integer in $[0, 100]$ that aims to be closest to two-thirds of the next group average ([PROMPT]).

Scientific workflow prediction (Sci. Workflow Pred.). Five generation sub-tasks (idea generation, method recommendation, outcome prediction, impact prediction, and title prediction), each predicting one workflow aspect from the preceding aspects of the same paper ([PROMPT]).

Economics contest (IEO). The model selects A/B/C/D for each multiple-choice problem from the International Economics Olympiad ([PROMPT]).

C Metrics Details

In tasks constructed from economic games, players' choice ranges differ across games (e.g., in the Dictator game, a player can choose a number between 0 and 100, but for the Public Goods game, a player can choose between 0 and 20). We normalize the players' choices into range $[0, 100]$. We then average these normalized W across all games. A lower W score indicates that the predicted distributions are closer to observed human behavior distributions.

Lower bound for W. Following Suh et al. (2025), we estimate the lower bound Wasserstein distance by bootstrapping the human respondents. For each task, we sample (with replacement) as many human responses as observed, and compute the Wasserstein distance between the observed human distribution and the bootstrapped distribution, averaging over evaluation units (a game, a survey item, or a personality dimension) in the same way as the model scores. We repeat this 1000 times and report the mean as the *lower bound for W* row of Table 3.

D Model Fine-tuning

`BehaviorBench` sets a higher bar for behavior foundation models, covering four capability categories and evaluating performance at both the individual and population levels. To test whether this bar can be met by behavioral fine-tuning rather than by scale alone, we fine-tune two open-weight LLMs on a broader set of behavioral tasks than prior behavioral fine-tuning efforts (Xie et al., 2025), with explicit coverage of diverse capabilities, behavioral contexts, and human populations.

D.1 Backbone Models and Fine-tuning Schema

For fine-tuning, we use two widely-used open-weight backbones at different scales: Qwen3-4B (Team, 2025) and Llama-3.3-70B (Team, 2024). Using two backbones an order of magnitude apart in size lets us separate the effect of behavioral fine-tuning from the effect of model capacity, and each fine-tuned model is compared directly against its own untuned backbone. We use parameter-efficient supervised fine-tuning, applying Low-Rank Adaptation (LoRA) (Hu et al., 2022) (rank 8, $\alpha = 32$) to all linear layers via the `ms-swift` framework (Zhao et al., 2025). Both backbones are trained for one epoch on 4×A100 80GB GPUs with a peak learning rate of 10^{-4}, sequence length 2048, and effective batch size 128. Fine-tuning the 4B and 70B backbones takes around 60 and 600 A100 hours respectively.

D.2 Fine-tuning Tasks and Data

Inspired by BehaviorBench, we formulate fine-tuning tasks as described in Section 2.2 and curate fine-tuning data from the three sources described in Section 2.1. Although fine-tuning shares the same task formulation with BehaviorBench, the fine-tuning data do not overlap with the BehaviorBench evaluation data. The splitting procedures are summarized in Appendix B.3.

Specifically, we instantiate 11 fine-tuning tasks organized by the general capabilities they target: behavior prediction and simulation, strategic decision-making, subject-trait inference, and the application of behavioral knowledge. The task mixture is deliberately broad. It exposes both backbones to all components in the framework of Equation 1 (i.e., behavior y, subject traits x, context c, and behavioral science knowledge $\mathbb{K}$). This breadth is intended to encourage generalization across tasks, contexts, and populations, rather than specialization in any single task or dataset. Details on fine-tuning data are described in Section D.3.

D.3 Fine-tuning Data Recipe

We instantiate eleven fine-tuning tasks that span the three data sources of Section B.1 and cover the four capability axes of BehaviorBench. All instances are drawn from the fine-tuning partitions defined in Section B.3, so no subject or paper that appears in the BehaviorBench evaluation set ever appears in fine-tuning. Per-task counts are listed in Table 5. We describe the tasks below grouped by the BehaviorBench capability they target.

Behavior prediction and simulation. Six tasks. Three on MobLab games: single-round behavior simulation, multi-round behavior prediction, and across-context single-round prediction. Three on the Big Five survey: response prediction from demographics, sequential response prediction in survey order, and masked response prediction over the full 50-item survey.

Subject-trait inference. Three Big Five tasks: dimension-score prediction from demographics, leave-one-dimension prediction across the five dimensions, and age prediction from the five dimension scores. The dimension-score-from-demographics task is upsampled by a factor of three in the final mix (three identical copies are shuffled together with the rest). This is because preliminary experiments show that this task requires more training to improve on the validation set.

Strategic decision-making. One task: strategic Beauty Contest play. The Beauty Contest fine-tuning data uses a multi-winner expansion: when several integers tie as winners against the recorded other players in a round, each winning integer becomes a separate fine-tuning example (capped at 10 per round).

Behavioral knowledge application. One task with five sub-tasks (idea generation, method recommendation, outcome prediction, impact prediction, and title prediction), each predicting one workflow aspect from the preceding aspects.

E Verbalized vs. Simulated Distributions

The distributional metrics in Table 3 score *simulated distributions*: a model is repeatedly prompted for individual information, and the aggregated response distribution is compared against the human response distribution. An alternative is to elicit a *verbalized distribution*: the model is asked a single question about the population and returns the full probability distribution as text, without any sampling (Meister et al., 2025). We evaluate the verbalized setting on the nine distributional tasks of BehaviorBench.

Table 6 reports the per-task W of the verbalized distributions. We have two main observations:

- **The model rankings under the two settings are positively correlated.** Average ranks over the nine tasks correlate at Spearman $\rho = 0.51$ ($p = 0.04$, Figure 3 Left). This means that the two settings largely measure a shared capability: models that simulate the population distribution well also tend to describe it well. The behavior foundation models are the main exception, e.g., Centaur-70B ranks second-best on simulated distributions but last on verbalized ones.

Model	Game Behav.	Multi-Round	Across-Ctx	Demo. To Resp.	Seq. Resp.	Masked Resp.	Demo. To Pers.	Across-Dim Pers.	Pers. To Demo.	Avg. Rank
Qwen3-4B	12.24	15.23	11.62	0.58	0.56	0.56	9.73	10.34	12.25	14.56
Llama3.3-70B	10.45	15.07	10.96	0.51	0.48	0.59	3.54	4.59	12.32	12.22
DeepSeek V3.2	9.40	10.90	9.02	0.50	0.48	0.51	5.16	4.94	11.21	10.56
GLM-5.2	6.68	8.03	6.47	0.40	0.41	0.42	3.98	3.51	9.79	6.44
Claude Sonnet 4.6	8.67	9.24	7.41	**0.36**	0.38	0.40	2.68	3.21	6.75	4.33
Claude Opus 4.6	9.69	9.63	8.14	0.36	**0.37**	0.42	2.02	2.38	10.89	5.67
Claude Opus 5	8.85	**6.34**	7.69	0.36	0.38	**0.39**	**1.25**	**1.36**	5.90	**2.56**
GPT 4.1	9.73	9.56	9.70	0.40	0.40	0.41	2.71	2.71	9.85	7.33
GPT 5.4	9.40	7.22	8.09	0.43	0.41	0.44	3.77	3.85	6.91	7.44
GPT 5.6 Terra	6.90	7.55	6.63	0.39	0.39	0.43	2.79	2.95	6.81	5.11
GPT 5.6 Sol	6.54	6.83	6.44	0.39	0.41	0.45	2.05	2.45	5.33	3.89
Gemini 3.1 Flash	9.67	11.31	9.47	0.42	0.39	0.45	5.02	5.41	12.37	10.22
Gemini 3.1 Pro	**5.36**	8.56	**4.21**	0.41	0.41	0.44	2.13	2.43	**1.86**	4.44
Socrates-Llama3-8B-DPO	9.44	14.78	10.25	0.73	0.70	0.72	10.08	10.35	15.39	14.56
Socrates-Qwen-14B-DPO	10.22	15.72	10.32	0.54	0.54	0.57	7.03	7.17	12.43	13.89
Centaur-70B	32.31	40.36	38.05	0.78	0.77	0.87	5.24	4.86	15.73	15.89
OSim-8B	11.89	17.05	13.78	0.53	0.53	0.62	8.06	8.33	9.70	13.89

Table 6: Direct distribution elicitation: Wasserstein distance (↓) between each model's verbalized distribution and the human distribution, per task. **Bold** and underline denote the best and second best performance. Task names follow Table 3.

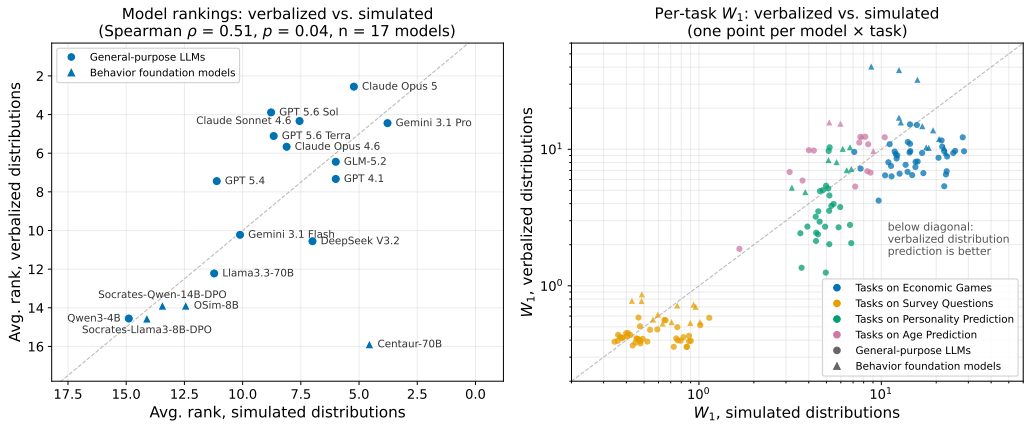


Figure 3: Verbalized vs. simulated distributions. **(Left)** Average model rank for verbalized and simulated distributions. **(Right)** Per-task Wasserstein distance under the two settings. One point denotes the performance of one model in one task, colored by task family. Points below the diagonal are cases where the verbalized distribution is more accurate than the simulated one.

- **Verbalized distributions tend to be more accurate than simulated ones.** Most general-purpose models attain a lower mean Wasserstein distance when verbalizing the distribution than when being sampled (Figure 3 Right). This aligns with the knowledge-to-simulation gap reported for opinion surveys by Meister et al. (2025): a model can often describe a population distribution more faithfully than it can sample from it. On the contrary, behavior foundation models tend to perform better with simulated distributions than verbalized distributions, which is different from general-purpose LLMs.

F Per-game Results for Behavior Prediction and Simulation

This appendix reports the per-game results for the three economic-game tasks summarized in the main results table (Table 3): single-round game behavior simulation, multi-round behavior prediction, and across-game single-round prediction. For each task we report distributional alignment with Wasserstein distance (W, lower is better) and individual-level prediction with mean absolute error (MAE, lower is better). The single-round simulation results appear in Table 7, the multi-round prediction results in Tables 8 and 9, and the across-game prediction results in Tables 10 and 11.

Model	Dictator	Proposer	Responder	Investor	Banker	Pub.Goods	Bomb	Beauty Contest	Push/Pull	Avg
Qwen3-4B	21.3	41.0	33.6	29.0	23.2	22.8	18.9	33.9	18.2	26.9
Llama3.3-70B	20.1	14.8	33.4	21.3	30.2	18.8	18.6	15.3	45.3	24.2
HY3	25.0	9.0	31.7	31.0	36.1	42.8	17.7	34.2	42.6	30.0
DeepSeek V3.2	**13.7**	8.0	7.6	27.1	18.9	22.9	17.7	18.0	18.0	16.9
GLM-5.2	19.0	7.6	14.8	20.3	19.3	20.2	17.7	17.0	47.7	20.4
Claude Sonnet 4.6	25.8	13.6	18.7	30.8	19.1	19.5	17.7	19.5	57.0	24.6
Claude Opus 4.6	25.8	13.3	23.8	33.3	19.2	23.6	17.7	20.4	57.0	26.0
Claude Opus 5	25.8	13.7	14.2	37.3	18.6	37.2	17.7	19.1	57.0	26.7
GPT 4.1	25.8	16.5	32.4	47.1	19.6	21.3	15.4	25.1	44.5	27.5
GPT 5.4	25.8	12.3	32.4	37.9	37.1	41.8	17.7	35.5	42.0	31.4
GPT 5.6 Terra	25.8	9.3	32.2	7.1	26.4	28.6	17.7	35.5	42.5	25.0
GPT 5.6 Sol	25.8	8.7	29.2	24.7	27.2	12.8	17.7	34.7	43.0	24.9
Gemini 3.1 Flash	25.8	12.0	14.9	57.8	22.8	53.7	17.7	22.5	54.0	31.3
Gemini 3.1 Pro	25.8	11.1	32.4	25.0	29.5	16.9	17.7	18.6	15.0	21.3
Socrates-Llama3-8B-SFT	26.6	7.3	16.1	**5.5**	**5.6**	**9.0**	9.6	**10.2**	57.0†	**16.3**
Socrates-Llama3-8B-DPO	25.8	12.0	18.1	33.6	22.9	25.1	17.6	25.3	57.0	26.4
Socrates-Qwen-14B-SFT	19.9	**4.4**	16.7	16.1	10.1	11.6	10.4	15.3	57.0†	17.9
Socrates-Qwen-14B-DPO	25.8	19.7	14.7	37.5	22.9	24.6	17.7	14.0	**3.6**	20.0
Centaur-70B	26.2	4.8	**5.7**	19.3	29.9	19.0	**5.6**	15.7	26.0	16.9
OSim-8B	22.8	7.4	32.6	37.5	16.7	18.3	13.5	17.3	30.7	21.9

Table 7: Single-round game behavior simulation: Wasserstein distance (↓). **Bold** and underline denote the best and second best performance. † The model produced no parseable Push/Pull response, so its Push/Pull W is set to the worst observed value in this column.

Model	Dictator	Investor	Banker_inv50	Banker_inv100	Pub.Goods	Bomb	Beauty Contest	Push/Pull	Avg
Qwen3-4B	13.5	22.2	19.3	7.7	15.0	13.6	9.2	4.0	13.1
Llama3.3-70B	21.8	12.6	27.3	13.8	15.0	12.0	7.3	12.8	15.3
HY3	2.2	14.8	**5.1**	3.7	42.5	16.0	9.0	33.0	15.8
DeepSeek V3.2	14.2	6.1	18.0	11.4	**5.5**	12.7	9.9	3.0	10.1
GLM-5.2	3.6	3.2	19.9	7.9	16.8	16.2	8.9	4.4	10.1
Claude Sonnet 4.6	5.9	8.3	16.9	9.4	17.5	17.2	8.7	**0.8**	10.6
Claude Opus 4.6	8.5	9.8	19.3	9.5	9.8	17.3	8.6	10.4	11.7
Claude Opus 5	3.2	11.6	18.1	9.0	11.5	16.8	9.6	12.6	11.5
GPT 4.1	2.5	**2.3**	15.1	8.4	8.4	5.0	8.1	5.6	**6.9**
GPT 5.4	2.1	17.3	9.7	**2.5**	34.7	17.2	9.4	28.6	15.2
GPT 5.6 Terra	2.0	18.4	7.3	2.6	22.3	17.3	9.2	32.8	14.0
GPT 5.6 Sol	4.2	22.5	8.2	3.8	32.5	17.3	9.0	31.2	16.1
Gemini 3.1 Flash	8.6	27.2	19.1	9.1	7.4	16.1	10.3	50.8	18.6
Gemini 3.1 Pro	**1.6**	14.7	10.9	7.1	24.4	16.4	9.7	15.0	12.5
Socrates-Llama3-8B-SFT	15.6	15.8	9.1	29.3	19.6	11.3	19.0	50.8†	21.3
Socrates-Llama3-8B-DPO	26.5	12.0	22.5	14.0	21.4	9.4	11.6	40.1	19.7
Socrates-Qwen-14B-SFT	18.5	5.5	6.9	11.5	12.8	6.8	14.0	50.8†	15.9
Socrates-Qwen-14B-DPO	14.5	15.2	20.5	14.1	10.8	7.8	**6.6**	26.0	14.4
Centaur-70B	15.3	2.6	12.3	8.1	11.5	5.0	6.8	27.0	11.1
OSim-8B	16.5	14.9	20.5	9.9	14.1	**3.6**	7.9	12.0	12.4

Table 8: Multi-round game behavior prediction: Wasserstein distance (↓). **Bold** and underline denote the best and second best performance. † The model produced no parseable Push/Pull response, so its Push/Pull W is set to the worst observed value in this column.

F.1 Economic Games

G Distributions of Model Outputs

This section visualizes the model-predicted output distribution alongside the test-set distribution. The plots cover single-round game behavior simulation (Figures 4 and 5), multi-round game behavior prediction (Figures 6 and 7), and across-game single-round prediction (Figures 8 and 9). The top row of each figure shows the test-set distribution, and the remaining rows show the per-model predicted distributions.

Model	Dictator	Investor	Banker_inv50	Banker_inv100	Pub.Goods	Bomb	Beauty Contest	Push/Pull	Avg
Qwen3-4B	18.7	30.0	23.5	17.0	20.9	17.8	11.4	34.0	21.7
Llama3.3-70B	23.8	35.4	28.8	18.4	26.6	18.5	11.1	53.6	27.0
HY3	10.1	33.1	17.5	15.6	42.9	17.6	11.5	35.4	23.0
DeepSeek V3.2	20.2	25.0	24.9	16.0	22.5	17.4	11.6	29.4	20.9
GLM-5.2	10.3	23.9	22.4	**12.0**	24.3	16.5	11.1	30.4	18.8
Claude Sonnet 4.6	11.8	22.7	20.7	13.2	24.5	17.2	11.0	28.8	18.7
Claude Opus 4.6	13.0	22.7	22.8	13.2	21.7	17.3	11.0	32.4	19.3
Claude Opus 5	9.7	26.4	21.5	12.8	20.8	17.0	11.2	**22.6**	**17.7**
GPT 4.1	10.1	**21.0**	21.3	15.0	**18.3**	**16.3**	10.6	32.4	18.1
GPT 5.4	**9.4**	33.0	16.9	13.7	36.4	17.2	11.2	31.4	21.1
GPT 5.6 Terra	10.3	33.9	18.9	16.2	29.1	17.3	11.2	36.8	21.7
GPT 5.6 Sol	11.0	35.8	**16.2**	14.2	35.6	17.3	11.0	34.8	22.0
Gemini 3.1 Flash	14.2	36.3	24.2	14.1	26.2	17.9	11.7	59.2	25.5
Gemini 3.1 Pro	9.7	31.3	19.2	13.7	29.7	16.8	11.5	23.4	19.4
Socrates-Llama3-8B-SFT	25.9	37.6	28.1	36.9	32.1	28.1	27.7	100.0†	39.5
Socrates-Llama3-8B-DPO	27.9	26.5	23.9	17.8	26.9	18.4	18.7	57.8	27.3
Socrates-Qwen-14B-SFT	27.9	33.0	26.1	27.7	26.5	22.5	23.7	100.0†	35.9
Socrates-Qwen-14B-DPO	18.4	26.4	25.1	17.0	19.5	16.5	**10.0**	30.8	20.5
Centaur-70B	19.8	23.5	26.6	22.6	24.7	19.0	15.5	34.6	23.3
OSim-8B	21.0	27.1	27.1	20.3	22.0	17.5	10.9	35.6	22.7

Table 9: Multi-round game behavior prediction: MAE (↓). **Bold** and underline denote the best and second best performance. † The model produced no parseable Push/Pull response; all such cases are counted as incorrect predictions.

Model	Dictator	Proposer	Responder	Investor	Banker	Pub.Goods	Bomb	Beauty Contest	Push/Pull	Avg
Qwen3-4B	16.3	21.7	28.4	27.1	19.7	22.0	17.7	18.3	**9.5**	20.1
Llama3.3-70B	14.5	9.5	10.4	20.7	12.4	16.7	16.8	14.0	18.7	14.9
HY3	9.2	9.1	6.8	10.0	6.4	19.9	16.6	15.7	32.3	14.0
DeepSeek V3.2	16.2	5.4	9.8	11.4	16.6	11.1	10.0	17.3	12.9	12.3
GLM-5.2	8.7	6.6	13.2	22.0	17.0	15.6	16.6	18.4	50.1	18.7
Claude Sonnet 4.6	9.9	9.6	12.1	26.0	14.8	14.4	17.3	19.0	50.7	19.3
Claude Opus 4.6	11.4	7.9	16.5	26.8	16.6	12.0	16.7	18.7	54.3	20.1
Claude Opus 5	8.1	7.5	11.6	21.0	15.0	12.0	14.6	18.9	29.7	15.4
GPT 4.1	15.4	9.4	11.0	16.6	16.0	18.4	10.3	18.3	30.9	16.3
GPT 5.4	5.9	8.4	5.2	10.4	6.0	14.9	17.4	21.1	37.2	14.0
GPT 5.6 Terra	11.2	10.0	4.0	8.2	6.8	14.1	17.6	29.0	33.1	14.9
GPT 5.6 Sol	12.4	8.3	**3.4**	12.5	5.7	**7.5**	17.2	16.9	36.3	13.4
Gemini 3.1 Flash	19.9	7.1	14.3	11.9	14.7	12.4	13.6	20.5	41.1	17.3
Gemini 3.1 Pro	**5.2**	7.4	8.9	9.7	8.0	7.9	13.5	16.7	15.3	**10.3**
Socrates-Llama3-8B-SFT	21.8	6.4	8.3	7.9	7.6	8.1	8.3	**8.7**	54.3†	14.6
Socrates-Llama3-8B-DPO	23.7	7.5	5.6	23.7	21.1	26.1	9.8	28.1	46.9	21.4
Socrates-Qwen-14B-SFT	20.8	4.8	8.8	**6.7**	**4.0**	11.1	9.0	13.1	54.3†	14.7
Socrates-Qwen-14B-DPO	20.1	3.5	11.8	31.6	19.5	21.4	17.7	15.3	18.8	17.7
Centaur-70B	21.4	**2.9**	7.6	11.4	14.4	17.2	**5.3**	19.8	24.8	13.9
OSim-8B	11.0	13.0	33.6	29.0	16.0	17.6	17.6	16.2	37.1	21.2

Table 10: Single-round game behavior prediction given observations from other games: Wasserstein distance (↓). **Bold** and underline denote the best and second best performance. † The model produced no parseable Push/Pull response, so its Push/Pull W is set to the worst observed value in this column.

Model	Dictator	Proposer	Responder	Investor	Banker	Pub.Goods	Bomb	Beauty Contest	Push/Pull	Avg
Qwen3-4B	22.0	24.4	29.9	**32.7**	26.2	24.7	**17.7**	26.1	47.3	27.9
Llama3.3-70B	24.4	15.0	18.3	36.7	26.9	25.7	18.2	23.4	48.8	26.4
HY3	24.3	14.7	16.8	39.5	26.5	35.3	18.1	22.6	46.4	27.1
DeepSeek V3.2	24.3	13.6	16.5	40.2	24.5	30.2	21.9	23.1	52.9	27.5
GLM-5.2	21.6	13.3	16.8	36.0	25.6	25.9	18.1	19.9	54.4	25.7
Claude Sonnet 4.6	20.7	13.1	15.7	35.9	22.8	26.5	17.9	19.8	55.2	25.3
Claude Opus 4.6	**19.4**	**12.3**	19.0	35.4	24.4	27.7	18.2	20.2	56.1	25.9
Claude Opus 5	21.2	**12.3**	15.6	36.5	23.4	33.4	18.2	19.9	51.9	25.8
GPT 4.1	22.6	15.4	15.4	39.3	23.9	29.2	26.2	20.2	51.5	27.1
GPT 5.4	22.5	14.1	14.7	37.2	25.5	32.5	17.9	25.9	44.7	26.1
GPT 5.6 Terra	23.8	15.3	**12.4**	40.9	24.6	34.7	**17.7**	30.4	45.1	27.2
GPT 5.6 Sol	23.8	14.1	14.1	38.7	24.5	30.6	17.9	22.0	**44.0**	25.5
Gemini 3.1 Flash	24.8	14.0	20.0	45.5	23.9	37.2	21.7	24.0	53.6	29.4
Gemini 3.1 Pro	23.8	14.1	17.9	39.1	25.3	32.2	19.3	21.6	47.1	26.7
Socrates-Llama3-8B-SFT	28.7	23.4	25.5	39.6	24.8	39.6	30.8	26.8	100.0†	37.7
Socrates-Llama3-8B-DPO	25.0	14.6	16.5	34.0	**21.7**	28.2	21.7	30.7	56.8	27.7
Socrates-Qwen-14B-SFT	29.7	22.8	24.3	42.2	26.4	30.1	27.8	26.6	100.0†	36.7
Socrates-Qwen-14B-DPO	23.7	16.0	16.5	34.4	22.9	**24.3**	**17.7**	**19.7**	48.1	**24.8**
Centaur-70B	26.0	16.3	15.8	37.5	30.3	34.8	29.5	27.9	45.6	29.3
OSim-8B	24.0	21.3	33.6	35.6	24.7	26.3	33.6	20.4	45.3	29.4

Table 11: Single-round game behavior prediction given observations from other games: MAE (↓). **Bold** and underline denote the best and second best performance. † The model produced no parseable Push/Pull response; all such cases are counted as incorrect predictions.

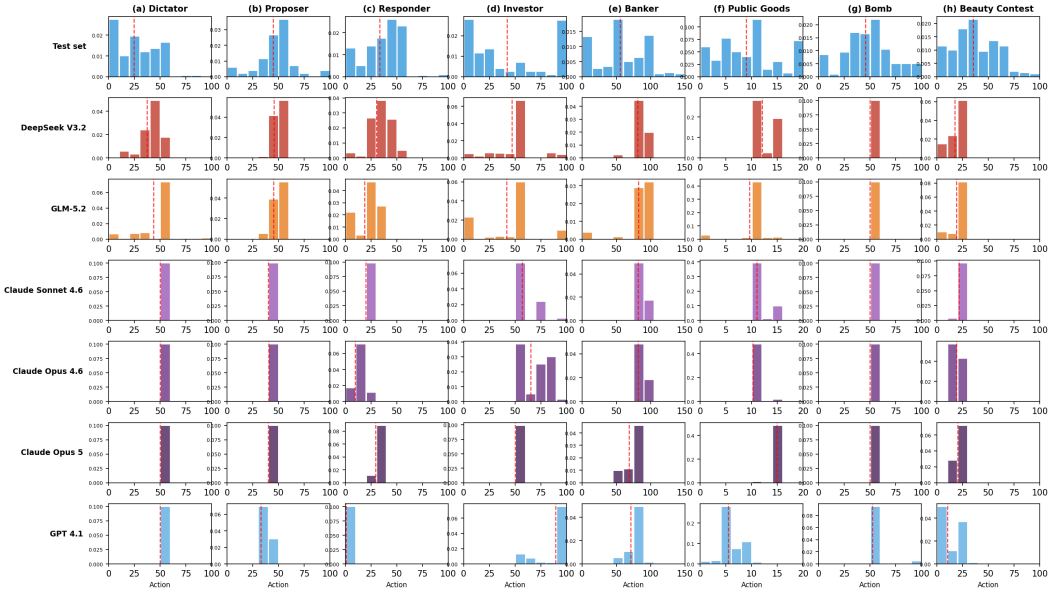


Figure 4: Distribution of model outputs in single-round game behavior simulation (Part 1).

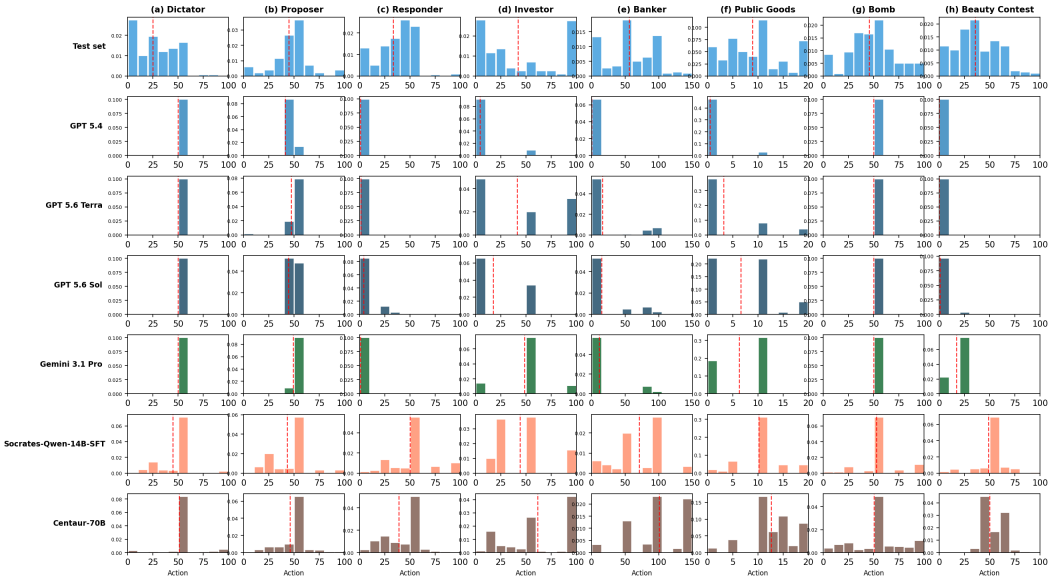


Figure 5: Distribution of model outputs in single-round game behavior simulation (Part 2).

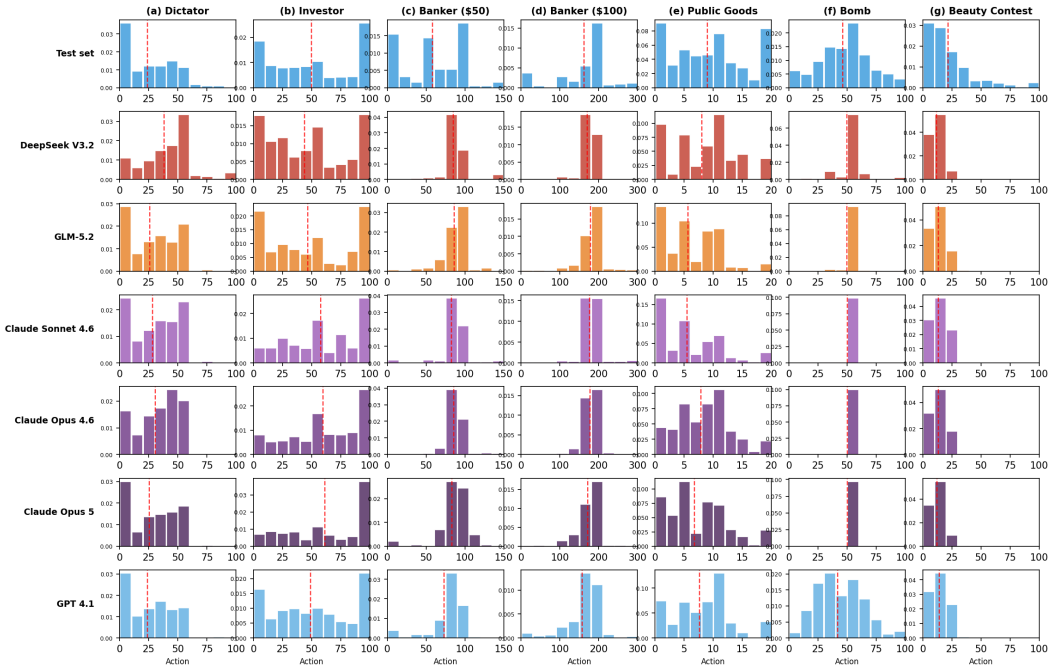


Figure 6: Distribution of model outputs in multi-round game behavior prediction (Part 1).

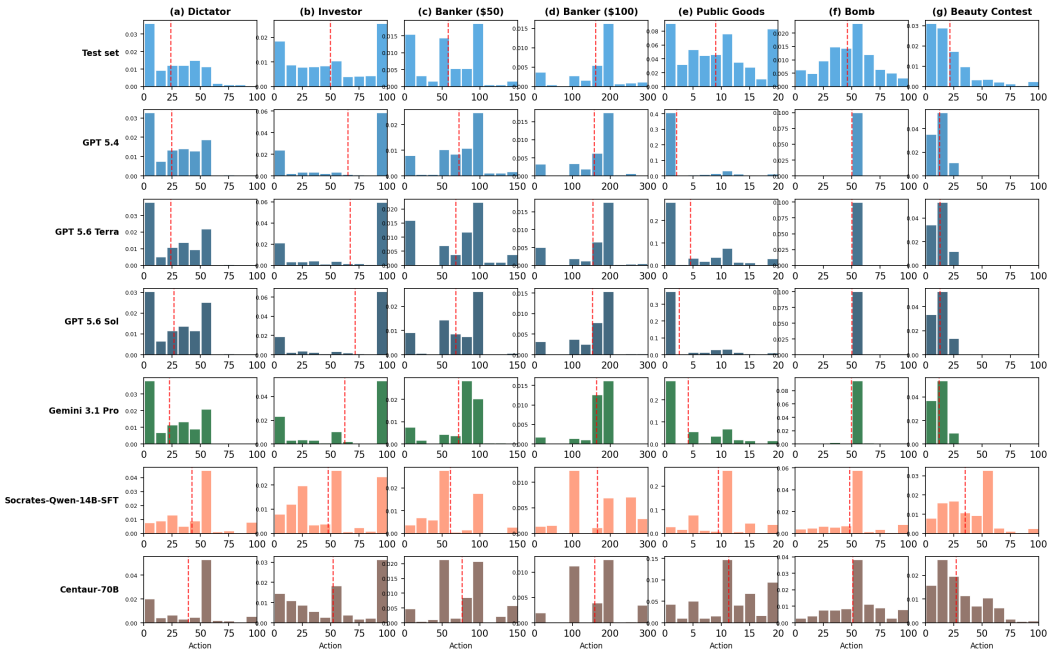


Figure 7: Distribution of model outputs in multi-round game behavior prediction (Part 2).

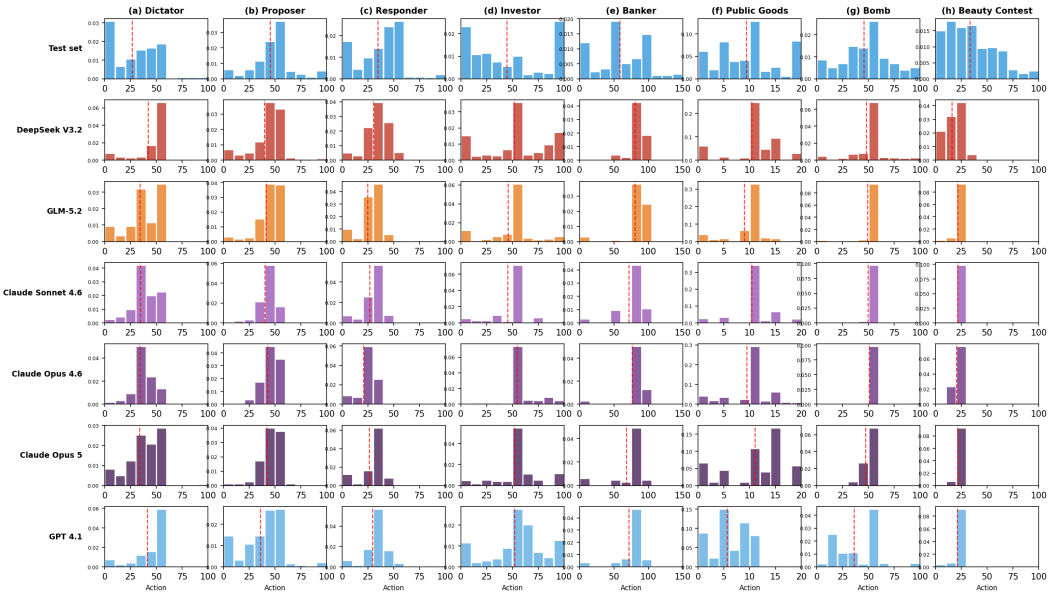


Figure 8: Distribution of model outputs in single-round game behavior prediction given observations from other games (Part 1).

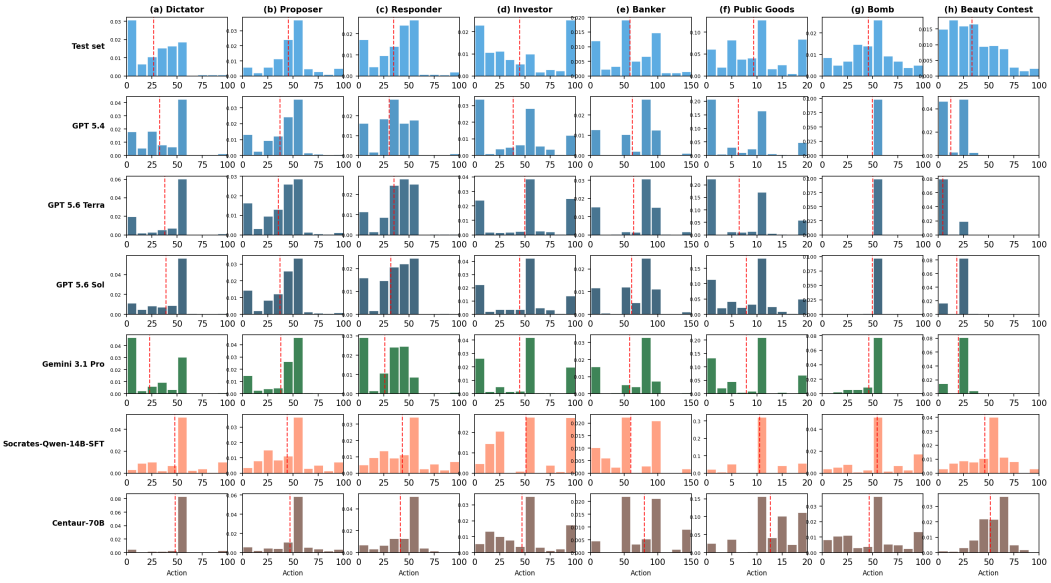


Figure 9: Distribution of model outputs in single-round game behavior prediction given observations from other games (Part 2).

H PROMPT TEMPLATES

This appendix documents all prompt templates used in the experiments. For each task, we show the system prompt and user prompt template. Placeholders are shown as {*variable*}. Section headers within prompts (originally in Markdown) are rendered in **bold**.

A. ECONOMIC GAMES — SINGLE-ROUND BEHAVIOR SIMULATION

All single-round economic game prompts share the following system prompt.

System prompt.

> You are a player who is playing an economics game.

Dictator Game.

> You are paired with another player. Your role is to decide how to divide $100 and the other player simply receives your choice. How would you like to divide the money? Please give only one concrete choice and highlight the amount you give to the other player in [] (such as [$x]).

Ultimatum Game — Proposer.

> This is a two-player game. You are the Proposer, and the other player is the Responder. As the proposer, you propose how to divide $100 and the Responder chooses either Accept or Reject. If accepted, the two of you will earn as described by the accepted proposal accordingly. If rejected, then both of you will earn $0. How much would you like to propose to give to the Responder? Please give only one concrete choice, and highlight the amount with [] (such as [$x]).

Ultimatum Game — Responder.

> This is a two-player game. You are the Responder, and the other player is the Proposer. The proposer proposes how to divide $100 and you, as the Responder, choose either Accept or Reject. If accepted, the two of you will earn as described by the accepted proposal accordingly. If rejected, then both of you will earn $0. As the Responder, what is the minimal amount in the proposal that you would accept? Please give only one concrete choice, and highlight the amount with [] (such as [$x]).

Trust Game — Investor.

> This is a two-player game. You are an Investor and the other player is a Banker. You have $100 to invest and you choose how much of your money to invest with the Banker. The amount you choose to invest will grow by 3x with the Banker. For example, if you invest $10, it will grow to $30 with the Banker. The Banker then decides how much of the money ($0-$30) to return to you, the Investor. How much of the $100 would you like to invest with the Banker? Please give only one concrete choice, and highlight the number with [] (such as [$x]).

Trust Game — Banker.

> This is a two-player game. You are a Banker and the other player is an Investor, and the goal for each player is to earn more. The Investor chooses how much of the money (up to $100) to invest with you. The amount the Investor invests will generate a 2x return with you (the current value of investment becomes 3x). To settle the investment, as the Banker, you get to decide how much of this total amount to return to the Investor and the rest will be kept as your profit. For example, you can choose to return $0 (therefore the investor will lose their investment), or you can return the entire 3x (initial investment + 2x profit) to the investor, or you can return any amount in between. Now, if the investor has invested $50 with you and the current value became $150, how much of the $150 would you like to return to the Investor? Please give only one concrete choice, and highlight the number with [] (such as [$x]).

Public Goods Game.

> In this public good game, you and 3 others will choose how much to contribute towards a water cleaning project. Each of you gets $20 per round to contribute between $0 and $20. The project has a 50% return rate. Your payoff relies on both your and others' contributions. Everyone benefits from the group's total

contribution. Your payoff in each round equals the amount you didn't contribute (endowment - contribution) plus everyone's benefit (total contributions * 50% return rate). Here are two examples to calculate your payoff.

Example one: You contributed $12; total group contributions were $20

Your Payoff = ($20-$12) + $20*50% = $18

Example two: You contributed $12; total group contributions were $30

Your Payoff = ($20-$12) + $30*50% = $23

We will play a total of 3 rounds, in the first round, how much of the $20 would you like to contribute? Please give a concrete number and highlight it with [] (e.g., [x]).

Bomb Risk Game.

There are 100 boxes, and one bomb has been randomly placed in 1 of 100 boxes. You can choose to open 0-100 boxes at the same time. If none of the boxes you open has the bomb, you earn points that are equal to the number of boxes you open. If one of the boxes you open has the bomb, you earn zero points. How many boxes would you open? Please give one concrete number and highlight it with [] (such as [x]).

Beauty Contest (Guessing) Game.

This is a number-guessing game where you will play with {*num_players*} other players. Each player chooses an integer between 0 and 100 which is hidden from other players. The player who chooses a number that is closest to 2/3 of the average wins the game. For example, if the average of the guesses is 60 and if you guess 40, i.e. 2/3 of 60, you win the game. What number would you like to guess? Please give a concrete number and highlight it with [] (e.g., [x]).

Push/Pull (Prisoner's Dilemma) Game.

You're paired with another player, each having a $400 'Push' card and a $300 'Pull' card. Your payoff depends on both players' card choices. Here are the scenarios:

- Both play 'Push': Each earns $400
- You play 'Push', the other player plays 'Pull': You earn $0, the other player earns $700
- You play 'Pull', the other player plays 'Push': You earn $700, the other player earns $0
- Both play 'Pull': Each earns $300

In 5 total rounds, which card would you play in the first round? Please give a concrete choice and highlight it with [] (i.e., [Push] or [Pull]).

Instruction suffix:
Important: Answer with [Push] or [Pull], not a number.

B. Economic Games — Multi-Round Behavior Prediction

The system prompt is the same as in Section A. Multi-round prompts predict a player's next action given their play history. The general user prompt template is:

General template.

{*game_instruction*}

Play History
In this session, you have already played {*num_rounds*} round(s). Here is the play history:
{*history*}

Your Choice
Now in Round {*next_round*}, what would you choose? Please give only one concrete choice and highlight the number with [] (such as [$x]).

Here {*game_instruction*} is the single-round prompt from Section A and {*history*} consists of per-round lines in one of the following formats.

History line formats.

> *Without other players (Dictator, Bomb):*
> - Round {*round_num*}. Your choice: [{*your_action*}].
>
> *With one other player — singular (Trust):*
> - Round {*round_num*}. Your choice: [{*your_action*}], the other player's choice: [{*others_action*}].
>
> *With multiple other players — plural (Public Goods):*
> - Round {*round_num*}. Your choice: [{*your_action*}], the other players' choices: [{*others_action*}].
>
> *Beauty Contest (custom):*
> - Round {*round_num*}. Your choice: [{*your_action*}], the group average is [{*group_avg*}], and 2/3 of that is [{*target*}]. {*round_outcome*}

Trust Banker — variable investment. In multi-round Trust Banker prompts, the investment amount varies per round. The game rules (same as Section A) are followed by the play history and a round-specific question:

> Now in this round, the investor has invested ${*investment*} with you and the current value became ${*total*}. How much of the ${*total*} would you like to return to the Investor? Please give only one concrete choice and highlight the number with [] (such as [$x]).

Beauty Contest — multi-round. The guessing game uses simplified rules (without the round count) and a modified question:

> Now you will play the game again (Round {*next_round*}). What number would you like to guess? Please give a concrete number and highlight it with [] (e.g., [x]).

Push/Pull — multi-round. The Push/Pull game uses [Push]/[Pull] choices instead of numeric values:

> Which card would you play in Round {*next_round*}? Please give a concrete choice and highlight it with [] (i.e., [Push] or [Pull]).
>
> Respond with only [Push] or [Pull]. Do not explain.

C. Economic Games — Across-context Single-round Prediction

The system prompt is the same as in Section A. Given a player's history from other games, the model predicts their choice in a new target game.

User prompt template.

> **Play History**
> You have played the following economics games before.
>
> {*history_section*}
>
> **Your Choice**
> Now you are playing a new game with the game instruction below:
>
> {*target_instruction*}

Each entry in {*history_section*} is formatted as:

> {*index*}. {*game_name*}
> - Game instruction: {*instruction*}
> - Your choice: [{*choice*}]

The {*target_instruction*} is the single-round prompt from Section A for the target game.

D. BIG FIVE PERSONALITY SURVEY

Task 1: Personality Score from Demographics.

System: You are an expert in psychology. Given a person's demographics, your task is to predict this person's BigFive dimensionality scores.

User:
Demographics
{*demographics*}

BigFive Dimensionality Scores
Each dimensionality score ranges from 10 to 50, with 10 indicating the lowest score in that dimension and 50 indicating the highest score.

Output Format
Based on this person's demographics, please estimate this person's personality score in the *{*dimension_name*}* dimension. Please output a single number in the range from 10 to 50, highlighted in [] (e.g., [x]).

{*demographics*} is formatted as: "A {*age*}-year-old {*gender*} from {*country*}. {*hand*}-handed. The race is {*race*}. The native language is {*engnat*}." {*dimension_name*} is one of: Extraversion, Neuroticism, Agreeableness, Conscientiousness, Openness.

Task 2: Survey Response Prediction Given Demographics.

System: You are a {*age*}-year-old {*gender*} from {*country*}. You are {*hand*}-handed. Your race is {*race*}. Your native language is {*engnat*}.

User: The following item was rated on a five-point scale where 1=Disagree, 2=Partially Disagree, 3=Neutral, 4=Partially Agree, 5=Agree. Please select how this statement describes you and highlight your answer in [](such as [1],[2],[3],[4],or [5]): {*question*} Only output your answer in brackets.

Task 3: Age Prediction from Personality.

System: You are an expert in psychology. Given a person's BigFive personality test dimensionality scores, your task is to predict this person's age.

User:
BigFive Dimensionality Scores
Each dimensionality score ranges from 10 to 50, with 10 indicating the lowest score in that dimension and 50 indicating the highest score.

- Openness: {*O_score*}
- Conscientiousness: {*C_score*}
- Extraversion: {*E_score*}
- Agreeableness: {*A_score*}
- Neuroticism: {*N_score*}

Output Format
Please predict this person's age. Please output a single number indicating the age highlighted with [] (e.g., [x]). Only output your answer in brackets.

Task 4: Sequential Survey Response Prediction.

System: You are an expert in psychology. Given a subject's previous answers in a Big Five personality test (in the order they answered them), your task is to predict this subject's answer to the next question.

User:
Subject's Previous Answers
The following items were rated on a five point scale where 1=Disagree, 2=Slightly Disagree, 3=Neutral, 4=Slightly Agree, 5=Agree.

{*previous_answers*}

The Next Question
{*next_question*}

Output Format
Please predict the subject's answer to the next question and highlight your prediction in [](such as [1],[2],[3],[4],or [5]). Only output your answer in brackets.

Items are numbered by original survey position (1–50); no item codes are shown.

Task 5: Masked Survey Response Prediction.

System: You are an expert in psychology. Given a subject's answers to 9 questions measuring the same Big Five personality dimension, your task is to predict this subject's answer to the remaining question in that dimension.

User:
Subject's Answers ({*dimension_name*} Dimension)
The following items were rated on a five point scale where 1=Disagree, 2=Slightly Disagree, 3=Neutral, 4=Slightly Agree, 5=Agree.

{*context_answers*}

The Remaining Question
{*target_question*}

Output Format
Please predict the subject's answer to the remaining question and highlight your prediction in [](such as [1],[2],[3],[4],or [5]). Only output your answer in brackets.

{*context_answers*} lists 9 items numbered 1–9 (no item codes).

Task 6: Across-Dimension Personality Prediction.

System: You are an expert in psychology. Given a person's BigFive personality test scores in four dimensions, your task is to predict the score in the remaining dimension.

User:
BigFive Dimensionality Scores
Each dimensionality score ranges from 10 to 50, with 10 indicating the lowest score in that dimension and 50 indicating the highest score.

{*dimension_scores*}

Output Format
Based on the four dimension scores above, please predict this person's score in the *{*target_name*}* dimension. Please output a single number in the range from 10 to 50, highlighted in [] (e.g., [x]). Only output your answer in brackets.

{*dimension_scores*} lists 4 of the 5 dimensions (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) with their scores.

E. Scientific Workflow Prediction

System prompt.

You are an expert in research tasked with generating detailed prompts for various aspects of academic research papers. Each task involves creating a specific type of prompt based on the provided information. Here are the definitions of each part you will work with:

- **Context:** The status quo of related literature or reality which motivated this study. This could normally be a problem, a research question, or a research gap that has not been successfully addressed by previous work. This is anything that happened before this study.
- **Key Idea:** The main intellectual merit of this paper, often in comparison to the context. This could normally be a novel idea or solution proposed in this paper that distinguishes it from what's already done in literature. This is proposed in this study.

- **Method:** The specific research method that investigates and validates the key idea. This could be an experimental setup, a theoretical framework, or other necessary methodology to implement and/or evaluate the key idea. This is performed in this study.
- **Outcome:** The factual statement about the study output. This could be the experiment results and any other measurable outcome that has occurred. It marks whether the key hypothesis is testified or not. This is produced in this study.
- **Projected Impact:** The author-anticipated impact of the work on the field, and potential further research identified by the author that may improve or extend this study. This is anything being anticipated but has not happened yet.

Task user prompt templates. The five workflow prediction tasks use the following user prompt templates:

1. Idea Generation:
Given the context: '{*context*}', generate one key idea that could advance this area of study. Output only the key idea without explanation.

2. Method Recommendation:
Given the context: '{*context*}' and the key idea: '{*key_idea*}', recommend one method to validate this idea. Output only the method without explanation.

3. Outcome Prediction:
Based on the context: '{*context*}', the key idea: '{*key_idea*}', and the method: '{*method*}', predict the outcome of this research. Output only the outcome without explanation.

4. Impact Prediction:
Based on the context: '{*context*}', the key idea: '{*key_idea*}', the method: '{*method*}', and the outcome: '{*outcome*}', suggest the projected impact for this research. Output only the impact without explanation.

5. Title Prediction:
Given the context: '{*context*}', the key idea: '{*key_idea*}', the method: '{*method*}', the outcome: '{*outcome*}', and the projected impact: '{*projected_impact*}', predict the title of this research paper. Output only the title without explanation.

F. IEO Economics Contest

System: You are an expert in economics. Answer the following multiple-choice question by selecting the correct option (A, B, C, or D). Provide only the letter of your answer.

User: {*question*}

A. {*option_A*}
B. {*option_B*}
C. {*option_C*}
D. {*option_D*}

G. Contextual Factor Inference — Dictator Game

System prompt.

You are an expert in behavioral economics.

User prompt template.

Given the observation of an experiment on the Dictator game, your task is to infer what experiment designs could lead to the observed treatment effect. In the Dictator game, given an endowment of money, one player (the dictator) chooses how much of the money to keep and how much to give to a second player.

In an economic experiment of the Dictator game, we observed the subject behaviors – the proportion of money to share – {*direction*} compared to the standard game design. Please list possible experiment designs that could lead to this treatment effect. Please rank output experiment designs by confidence.

{*direction*} is one of: *increased*, *decreased*.

I LICENSES FOR EXISTING ASSETS

BehaviorBench is constructed from the publicly available data sources described in Section B.1. We summarize the licensing and terms-of-use status of each below.

MobLab economic-games data. The MobLab game logs we use are released by Mei et al. (2024) via the public repository https://github.com/yutxie/ChatGPT-Behavioral under the MIT License.

Big Five Personality Test. The Big Five responses are obtained from the publicly available Kaggle dataset https://www.kaggle.com/datasets/lucasgreenwell/ocean-five-factor-personality-test-responses, which mirrors data distributed by the Open-Source Psychometrics Project (https://openpsychometrics.org/_rawdata/). The dataset is released under the Database Contents License (DbCL) v1.0 (https://opendatacommons.org/licenses/dbcl/1-0/).

Scientific articles (AER, NHB). We use only the titles and abstracts of articles published in the *American Economic Review* (American Economic Association) and *Nature Human Behaviour* (Springer Nature). All articles are properly cited, and titles and abstracts are used for non-commercial research purposes consistent with standard fair-use practice for bibliographic data.

International Economics Olympiad. Past IEO multiple-choice problems are publicly distributed for educational use (https://ieo-official.org/). We use them only for non-commercial research evaluation.