

# 000 SIMBENCH: BENCHMARKING THE ABILITY OF LARGE 001 LANGUAGE MODELS TO SIMULATE HUMAN BEHAV- 002 IORS 003 004

005 **Anonymous authors**  
006  
007 Paper under double-blind review  
008  
009  
010  
011  
012

## ABSTRACT

013 **Changes are marked in blue.** Large language model (LLM) simulations of human  
014 behavior have the potential to revolutionize the social and behavioral sciences, *if*  
015 *and only if* they faithfully reflect real human behaviors. Current evaluations are  
016 fragmented, based on bespoke tasks and metrics, creating a patchwork of incom-  
017 parable results. To address this, we introduce SIMBENCH, the first large-scale,  
018 standardized benchmark for a robust, reproducible science of LLM simulation.  
019 By unifying 20 diverse datasets covering tasks from moral decision-making to  
020 economic choice across a large global participant pool, SIMBENCH provides the  
021 necessary foundation to ask fundamental questions about when, how, and why  
022 LLM simulations succeed or fail. We show that, while even the best LLMs today  
023 have limited simulation ability (score: 40.80/100), performance scales log-linearly  
024 with model size. Simulation performance is not improved by increased inference-  
025 time compute. We discover an alignment-simulation trade-off: instruction-tuning  
026 improves performance on low-entropy (consensus) questions but degrades it on  
027 high-entropy (diverse) ones. Models particularly struggle when simulating specific  
028 demographic groups. Finally, we demonstrate that simulation ability correlates  
029 most strongly with deep, knowledge-intensive reasoning (MMLU-Pro,  $r = 0.939$ ).  
030 By making progress measurable, we aim to accelerate the development of more  
031 faithful LLM simulators.

032 We combine **20 datasets**  
033 in a unified format.  
034

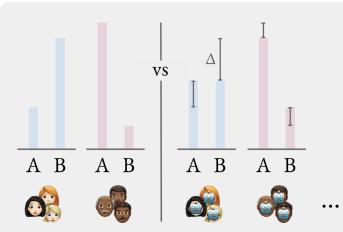
035 ChaosNLI MoralMachineC  
036 Choices13k AfroBarometer  
037 OpinionQA OSPsychBig5  
038 NumberGame DICES990  
039 WisdomOfCrowds Jester  
040 LatinoBarometro ISSP ...  
041

A train will kill 5 people on the track. You can flip a switch to divert the train to a side track where it will kill just 2 people.

**What do you do?**  
A: Flip the switch  
B: Do nothing

Each dataset contains  
**multiple-choice questions**.

We test the ability of LLMs to simulate **group-level responses**.



043 Figure 1: SIMBENCH is the first large-scale benchmark to evaluate how well LLMs can simulate  
044 group-level human behavior across diverse simulation settings and tasks.  
045  
046

## 1 INTRODUCTION

049 Large-scale human experiments and surveys have long been essential tools for informing public  
050 policy, commercial decisions, and academic research. Running experiments and surveys, however, is  
051 costly and time-consuming. Large language models (LLMs) can potentially address this challenge by  
052 simulating human behaviors quickly and at low cost, to complement or even substitute human studies.  
053 This prospect, alongside encouraging early evidence on the efficacy of LLMs as simulators (Aher et al., 2023; Argyle et al., 2023; Horton, 2023), has motivated a large body of recent work across many

054 disciplines investigating the ability of LLMs to simulate human behaviors (Binz et al., 2025; Bisbee  
 055 et al., 2024; Dominguez-Olmedo et al., 2024; Manning et al., 2024; Hu & Collier, 2025, *inter alia*).  
 056

057 However, this rapid exploration has produced a fragmented body of evidence. Most studies evaluate  
 058 a narrow set of LLMs on a specific task, yielding varied and sometimes contradictory results that  
 059 make it difficult to draw broader conclusions (§5). The field lacks a unified framework to determine  
 060 when, how, and why LLM simulations succeed, or how to train better simulators.

061 To confront this challenge and move LLM simulation from a collection of isolated studies to a  
 062 robust, reproducible science, we introduce **SIMBENCH**: the first large-scale, standardized benchmark  
 063 for human behavior simulation. By harmonizing 20 diverse datasets – spanning moral dilemmas,  
 064 economic games, and psychological assessments across a vast global participant pool – **SIMBENCH**  
 065 provides the essential instrument to rigorously measure and compare simulation fidelity across models,  
 066 tasks, and populations.

067 Using this benchmark, we move beyond isolated experiments to build a comprehensive picture of  
 068 LLM simulation. We structure our investigation through six research questions. We first establish a  
 069 baseline, asking **how well current LLMs perform** (RQ1) and **how characteristics like model size**  
 070 **and inference-time compute affect their ability** (RQ2). We find that even top models struggle (top  
 071 score: 40.8/100), though performance scales log-linearly with size and is surprisingly not improved  
 072 by scaling inference-time compute. Next, we explore the sources of this variance, asking **how task**  
 073 **selection** (RQ3) and **human response plurality** (RQ4) affect fidelity. Fidelity varies substantially  
 074 by task, and we uncover a key alignment-simulation tradeoff: instruction tuning’s mode-seeking  
 075 objective systematically improves performance on low-entropy (consensus) questions but actively  
 076 harms performance on high-entropy (pluralistic) ones. A causal analysis confirms this is the net  
 077 result of a beneficial instruction-following effect versus a harmful entropy-reduction effect. Finally,  
 078 we investigate practical implications, asking if LLMs are better at **simulating some demographic**  
 079 **groups than others** (RQ5) and **to what extent simulation ability correlates with other model**  
 080 **capabilities** (RQ6). We show models struggle most with religious/ideological groups and that  
 081 simulation ability correlates most strongly with deep, knowledge-intensive reasoning (MMLU-Pro,  
 $r = 0.939$ ).

082 Progress in AI is only possible through rigorous evaluation, and large-scale benchmarks such as  
 083 MMLU (Hendrycks et al., 2021) have significantly contributed to improvements in LLM capabilities.  
 084 In the same spirit, **SIMBENCH** provides the foundational infrastructure needed to move LLM  
 085 simulation from a collection of ad-hoc studies to a measurable and systematic science. **We note that**  
 086 **while predicting group-level response distributions is the instrument used here for benchmarking,**  
 087 **we view this as a foundational proxy for the broader capability of human behavior simulation.** We  
 088 acknowledge that full human simulation inherently includes interactive and complex dynamics not  
 089 fully captured by static response distributions (see Appendix A). All of **SIMBENCH** is available on  
 090 GitHub and HuggingFace (also in the supplementary material).

## 092 2 CREATING **SIMBENCH**

### 094 2.1 DATA CURATION

096 To create **SIMBENCH**, we combine a repository-driven approach, where we query major social  
 097 and behavioral science repositories (e.g., Harvard Dataverse, ICPSR, OSF), and a literature-driven  
 098 approach, where we identify key papers in relevant fields and trace back to their underlying data  
 099 sources. We then apply a strict set of selection criteria to all candidate datasets: **large participant**  
 100 **counts**, to ensure meaningful group-level distributions; **permissive licensing** to allow for redistribution;  
 101 **single-turn, self-contained questions**, to establish a standardized evaluation paradigm free  
 102 from multi-turn or contingent interactions; **multiple-choice or ordinal response formats**, to enable  
 103 quantitative evaluation; and **English-language questions** or validated translations for consistency.

104 These criteria are complemented by a curation strategy that balances competing objectives. We  
 105 prioritize **novelty**, favoring datasets not previously used in LLM simulation evaluation, but also  
 106 ensure **backward comparability** by including well-established benchmarks (e.g., OpinionQA,  
 107 ChaosNLI). Furthermore, we prioritize datasets with rich **sociodemographic data** to enable fine-  
 108 grained analysis of specific subpopulations (§2.4). However, we make targeted exceptions for datasets

108 like Jester and Choices13k, which, despite lacking demographic data, provide unique and essential  
 109 task diversity.  
 110

111 This principled selection process yields the **20** datasets that comprise SIMBENCH, which we list in  
 112 Appendix L, providing details on participants and example questions. To demonstrate the rigor of  
 113 our curation process and to serve as a resource for the community, we provide a list of datasets that  
 114 were considered but ultimately excluded in Appendix B. Crucially, SIMBENCH is fully modular by  
 115 design, so that future work can easily add more datasets using the processing pipeline described in  
 116 §2.3 below.  
 117

## 2.2 BENCHMARK PROPERTIES

119 The principled curation process results in a benchmark defined by two key properties: task diversity  
 120 and participant diversity.  
 121

122 **1) Task Diversity:** SIMBENCH covers a wide range of different tasks regarding the human behavior  
 123 they measure. SIMBENCH includes **decision-making** questions (e.g., in Choices13k, MoralMachine),  
 124 where participants are presented with a set of actions that concern themselves, and they have to  
 125 select the action they would hypothetically take. SIMBENCH also includes **self-assessment** questions  
 126 (e.g., in OpinionQA, OSPsychBig5), where participants are presented with a set of descriptions  
 127 or attributes, and they have to select the one that best describes themselves. Further, SIMBENCH  
 128 includes **judgment** questions (e.g., in ChaosNLI and Jester) where participants are presented with  
 129 some external object and a choice of labels, and they have to select the label they think fits best. Lastly,  
 130 SIMBENCH includes **problem-solving** questions (e.g., in WisdomOfCrowds and OSPsychMGKT),  
 131 where participants are presented with a set of answers to a factual question, and they have to select  
 132 the answer they think is correct. Consequently, LLMs have to accurately simulate several distinct  
 133 types of human behavior in order to perform well on SIMBENCH.  
 134

135 **2) Participant Diversity:** SIMBENCH captures a rich demographic landscape spanning more than 130  
 136 different countries across six continents (see Appendix O for a full country-level breakdown). While  
 137 five datasets include US-based crowdworkers, SIMBENCH prioritizes international representation:  
 138 samples from the Anglosphere West constitute only 27.9% of the data.<sup>1</sup> This substantial global  
 139 scope is driven by a diverse collection of sources: 3 datasets (e.g., LatinoBarometro, AfroBarometer)  
 140 exclusively feature participants from regions outside the US, 4 datasets (e.g., GlobalOpinionQA,  
 141 TISP) draw from multi-country samples across different continents, and 2 datasets collect responses  
 142 from a global pool of internet users. Importantly, 8 out of the 20 datasets employ representative  
 143 sampling techniques, enhancing the ecological validity of these constituent components. To perform  
 144 well on SimBench, LLMs must therefore demonstrate the ability to accurately simulate the behavior  
 145 of human participants across diverse cultural, linguistic, and socioeconomic backgrounds.<sup>2</sup>  
 146

## 2.3 UNIFYING SIMBENCH DATASET FORMATS

147 A core contribution of SIMBENCH is the harmonization of 20 heterogeneous datasets into a standard-  
 148 ized format. This process ensures that LLM performance can be compared rigorously across diverse  
 149 tasks and populations.  
 150

151 **Question Normalization:** We standardize all items into a multiple-choice format. Our normalization  
 152 is minimal, primarily consisting of mapping existing discrete options to standardized letter keys to  
 153 preserve the original experimental structure. For the few datasets (Jester) with continuous scales,  
 154 we map responses to discrete bins. We further ensure consistency by collapsing answer options  
 155 where appropriate, limiting the maximum to 26 choices (though typically fewer than 10), and using  
 156 the official English-language versions of all questions.<sup>3</sup> This is a deliberate choice to standardize  
 157 the evaluation and avoid confounding simulation ability with multilingual performance capabilities,  
 158

159 <sup>1</sup>We define the Anglosphere West as the U.S., Canada, U.K., Australia, and New Zealand. Even using a  
 160 broader definition of “the West” that includes Western Europe, these nations account for less than half (45.9%)  
 161 of the benchmark.  
 162

163 <sup>2</sup>Note that, while some constituent datasets recruit representative samples, SIMBENCH as a whole is not fully  
 164 representative of any single population. We discuss this limitation in Appendix A.  
 165

166 <sup>3</sup>We note that simulation ability may plausibly be correlated with prompt language, and encourage future  
 167 work in this direction.  
 168

162 ensuring that differences in scores reflect simulation fidelity rather than translation quality, even for  
 163 datasets originally collected in local languages.

164 **Response Aggregation:** To evaluate group-level simulation, we standardize all data into group-  
 165 level probability distributions. For the majority of our datasets, which provide raw individual-level  
 166 responses, we create these distributions by aggregating the data ourselves. Post-stratification weights  
 167 are applied whenever applicable (e.g. ESS). For the few datasets that are already provided in an  
 168 aggregated format (e.g., GlobalOpinionQA), we process and normalize their existing statistics to  
 169 conform to our benchmark’s schema.

170 We create the simulation targets in two ways: 1) **Default Grouping.** For every question in a dataset,  
 171 we create a baseline target by aggregating responses from all participants. This represents the “default”  
 172 population for that dataset (e.g., “US-based Amazon Mechanical Turk workers”) and is used to  
 173 measure general simulation ability. 2) **Specific Grouping:** For datasets with rich sociodemographic  
 174 data, we create more fine-grained targets by aggregating responses from participants sharing a specific  
 175 attribute (e.g., age or gender). These targets are essential for evaluating a model’s ability to simulate  
 176 narrower, more specific demographic groups. The available grouping variables for each dataset are  
 177 detailed in Appendix L.

178 Each simulation target is paired with a prompt that describes the corresponding group. This entire  
 179 harmonization process yields **10,930,271** unique question-group simulation targets. From this  
 180 comprehensive set, we curate our final benchmark splits (§2.4) to enable robust evaluation of LLM  
 181 simulation capabilities. **We note that while training data contamination remains an inherent risk, our**  
 182 **zero-shot aggregation task largely mitigates this by testing distributional prediction (see Appendix A**  
 183 **for a full discussion).**

#### 185 2.4 SIMBENCH SPLITS

187 The full set of over 10 million simulation targets is too vast for practical evaluation. We therefore  
 188 curate two distinct benchmark splits, each designed to probe a different facet of an LLM’s simulation  
 189 capabilities.

190 1) The **SimBenchPop** split covers all questions in all 20 datasets after processing as in §2.3. We  
 191 combine each question with the dataset-specific default grouping prompt to create one unique test case,  
 192 resulting in 7,167 test cases. We obtain the response distribution for each test case by aggregating all  
 193 individual responses to that test case over all participants in that dataset. Conceptually, **SimBenchPop**  
 194 **measures the ability of LLMs to simulate responses of broad and diverse human populations.**

195 2) The **SimBenchGrouped** split contains only the five large-scale survey datasets in SIMBENCH  
 196 (AfroBarometer, ESS, ISSP, LatinoBarometro, and OpinionQA) because for these datasets we have  
 197 enough participants to obtain meaningful group sizes even when selecting on a specific group attribute  
 198 (e.g., age = 30-49). For each dataset, we select questions that exhibit significant variation across  
 199 demographic groups, ensuring that the benchmark captures meaningful demographic differences  
 200 in responses. This results in 6,343 test cases overall. For more details on the sampling process,  
 201 see Appendix C. Conceptually, **SimBenchGrouped measures the ability of LLMs to simulate**  
 202 **responses from narrower participant groups based on specified group characteristics.**<sup>4</sup>

### 204 3 EXPERIMENTAL SETUP

206 **Tested Models:** To demonstrate the utility of SIMBENCH and answer our six research questions (§1),  
 207 we evaluate 45 state-of-the-art LLMs on SimBench. This includes both commercial and open-weight,  
 208 base and instruction-tuned models, with model sizes ranging from 0.5B to 405B parameters. Table 1  
 209 shows the full list of models.

210 **Model Elicitation:** For each model, we collect predictions for the two main splits of SIMBENCH  
 211 (§2.4). To obtain model response distributions, we use one of two methods, depending on model

213 <sup>4</sup>Ideally, we would also like to measure LLM simulation ability for intersectional groups that combine  
 214 multiple characteristics (e.g., female + age 30-49). However, selecting on multiple characteristics substantially  
 215 decreases group size, thus increasing sampling noise in the response distributions. Reliable evaluation of  
 intersectional group simulation ability would require datasets with more participants than we have access to.

216 type: 1) For base models, we directly extract **token probabilities** for each response option based  
 217 on first-token logits. This is a natural way of eliciting a distribution out of an LLM, especially a  
 218 base LLM. 2) For instruction-tuned models, we follow recent literature on LLM calibration and  
 219 distribution prediction (Tian et al., 2023; Meister et al., 2025) and use **verbalized distributions**,  
 220 e.g., “Option A: 30%, Option B: 70%”, elicited through prompting. We empirically validate this  
 221 methodological choice in Appendix E, which provides strong evidence that verbalized distributions  
 222 substantially and consistently outperform direct token probabilities for instruction-tuned models.  
 223 This ensures each model class is evaluated under its optimal conditions. For implementation details  
 224 and prompt formats, see Appendix D.

225 **Evaluation Metric:** To measure LLM simulation ability, we derive the **SIMBENCH** score  $S$  from the  
 226 Total Variation Distance (TVD). Conceptually,  $S$  quantifies the improvement of a model’s prediction  
 227  $Q$  over a uniform baseline  $U$ , relative to the human ground truth  $P$ :

$$228 \quad S(P, Q) = 100 \left( 1 - \frac{\text{TVD}(P, Q)}{\text{TVD}(P, U)} \right) \quad (1)$$

231 where a score of 100 indicates perfect alignment and 0 indicates performance equivalent to random  
 232 guessing. In practice, however, a direct point-wise calculation is undefined when the human distribu-  
 233 tion is uniform ( $P = U$ ). To ensure numerical stability, we compute the score  $S_i$  for each specific  
 234 test case  $i$  by normalizing against the *dataset-level* mean baseline:

$$236 \quad S_i = 100 \left( 1 - \frac{\text{TVD}(P_i, Q_i)}{\frac{1}{|D|} \sum_{j \in D} \text{TVD}(P_j, U_j)} \right) \quad (2)$$

238 where the denominator represents the average TVD between human responses and the uniform  
 239 distribution across all test cases  $j$  in the dataset  $D$ . This ensures the metric remains robust across  
 240 datasets with varying entropy. The final reported score for a model is the average of  $S_i$  across all  
 241 evaluated test cases.

## 243 4 RESULTS

### 245 4.1 RQ1: GENERAL SIMULATION ABILITY OF LLMs

247 To evaluate the general simulation ability of  
 248 LLMs, we measure their overall **SIMBENCH**  
 249 score  $S$  averaged across the two main splits of  
 250 **SIMBENCH** (Table 1 and Appendix Table 7).

251 We find that **even leading LLMs struggle to**  
 252 **simulate group-level human behaviors with**  
 253 **high accuracy**, as measured across the 20  
 254 datasets in **SIMBENCH**. Claude-3.7-Sonnet is  
 255 the best-performing model overall, but only  
 256 achieves a score of 40.80 out of a maximum  
 257 of 100 on **SIMBENCH**. This score indicates that  
 258 the response distributions predicted by Claude-  
 259 3.7-Sonnet are, on average, closer to a uniform  
 260 response distribution than to the true human re-  
 261 sponse distribution. The top open-weight model,  
 262 DeepSeek-R1, scores 34.52. The majority of the  
 263 45 models we test perform substantially worse  
 264 still, scoring less than 20. Notably, nine mod-  
 265 els we test score below 0, indicating that their  
 266 predicted response distributions are, on average,  
 267 even further away from the true human response  
 268 distribution than a uniform response distribution.  
 269 A comprehensive statistical analysis, detailed in  
 Appendix H, confirms that the performance dif-  
 ferences between most top-ranked models as well as within each model family are statistically

251 **Table 1: Overall simulation ability of represen-  
 252 tative LLMs** as measured by **SIMBENCH** score  $S$   
 253 averaged across the two main splits of **SIMBENCH**.  
 254 Reasoning models are highlighted in *italics*. A full  
 255 table with all 45 models is in Appendix Table 7.

Model	Type	Release	$S (\uparrow)$
<b><i>Top-Performing Models</i></b>			
Claude-3.7-Sonnet	Instr.	Closed	40.80
<i>Claude-3.7-Sonnet-4000</i>	Instr.	Closed	39.46
GPT-4.1	Instr.	Closed	34.55
<i>DeepSeek-R1</i>	Instr.	Open	34.52
Llama-3.1-405B-Instruct	Instr.	Open	28.40
<i>o4-mini-high</i>	Instr.	Closed	28.99
Qwen2.5-72B-Instruct	Instr.	Open	27.61
Qwen2.5-32B-Instruct	Instr.	Open	23.76
OLMo-2-32B-DPO	Instr.	Open	19.80
<b><i>Top-Performing Base Models</i></b>			
OLMo-2-32B	Base	Open	15.90
OLMo-2-13B	Base	Open	13.83
Qwen2.5-72B	Base	Open	13.34
Qwen2.5-32B	Base	Open	12.27
<b><i>Models Performing Below Uniform Baseline</i></b>			
Gemma-3-4B-PT	Base	Open	-0.65
Qwen2.5-3B-Instruct	Instr.	Open	-12.04
OLMo-2-7B-Instruct	Instr.	Open	-21.36

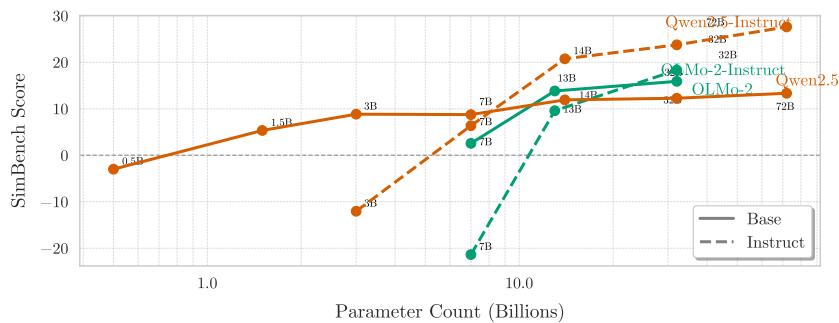
270 significant. Overall, these results consolidate the mixed findings from prior work into a clearer,  
 271 if somewhat sobering, picture. When evaluated across a diverse range of tasks and populations,  
 272 today’s LLMs are still far from being consistently reliable, general-purpose simulators. The stark  
 273 performance differences between models also caution strongly against the use of smaller, less capable  
 274 models for simulation, many of which perform worse than a simple uniform baseline.

#### 276 4.2 RQ2: IMPACT OF LLM CHARACTERISTICS ON SIMULATION ABILITY

278 While even the best models struggle to perform well on SIMBENCH, Table 1 also shows clear  
 279 differences across models. Therefore, we investigate how performance varies depending on model  
 280 characteristics, specifically 1) model size, and 2) test-time compute.

281 **1) Model Size** To evaluate the impact of model size on simulation ability, we plot SIMBENCH  
 282 Score  $S$  against model parameter count for the four LLM families that we can test across multiple  
 283 model sizes (Figure 2 and Appendix Figure 6). Our results suggest that [there is a log-linear scaling](#)  
 284 trend for LLM simulation ability. Across examined model families, for both base and instruction  
 285 models, an increase in parameter count generally corresponds to an increase in SIMBENCH score  $S$ ,  
 286 indicating better alignment between predicted and human response distributions. This relationship is  
 287 robustly supported by model families with comprehensive size variants (e.g., Qwen2.5, Llama-3.1),  
 288 though additional data points would be needed to [fully characterize the trajectory for families with](#)  
 289 [fewer models](#) (e.g., OLMo). There is also an interaction between model size and the effect of  
 290 instruction-tuning. While instruction-tuned models consistently outperform their base counterparts  
 291 at larger scales ( $>10$ B parameters), this relationship appears to invert for smaller models. For  
 292 example, the OLMo-2 base models outperform their instruction-tuned variants at the 7B and 13B  
 293 scale. Furthermore, the plot shows that **instruction-tuned models not only reach a higher peak**  
 294 **performance but also appear to scale more effectively**. The steeper slope of the dashed lines (e.g.,  
 295 for Qwen2.5-Instruct) compared to the solid lines suggests that instruction-tuning may improve a  
 296 model’s ability to capitalize on increases in parameter count for the simulation task. We present a  
 297 more comprehensive plot including all evaluated model families in [Appendix G](#), which confirms this  
 298 trend holds across considered models.

299 Overall, the clear positive scaling trends across model families suggest that, while simulation remains  
 300 a challenging task for even the best models today, further model scaling may well lead to highly  
 301 accurate LLM simulators in the future.



313 **Figure 2: Model parameter count vs. simulation ability.** We measure model size by parameter count  
 314 and simulation ability by SIMBENCH score  $S$  averaged across the two main splits of SIMBENCH.

315 **2) Test-Time Compute** To analyze the effects of increasing test-time compute on LLM simulation  
 316 ability, we conduct two sets of experiments. We compare the performance of two distinct o4-mini  
 317 checkpoints (‘low’ vs. ‘high’, which vary in the amount of reasoning efforts), and we assess Claude-  
 318 3.7-Sonnet with and without a 4000-token reasoning budget. Additionally, we apply a zero-shot  
 319 Chain-of-Thought (CoT) prompting strategy (Wei et al., 2022) to GPT-4.1 and DeepSeek-V3-0324  
 320 (see Appendix D for prompt details).

321 Our results suggest that **increasing test-time compute provides no meaningful benefit for LLM**  
 322 **simulation ability.** The o4-mini model has a minor improvement ( $S$  score: 28.20 → 29.54), while

Claude-3.7-Sonnet's performance slightly decreases ( $40.51 \rightarrow 39.46$ ). Similarly, applying CoT prompting leads to a small performance drop for GPT-4.1 ( $34.90 \rightarrow 33.11$ ) and a negligible change for DeepSeek-V3-0324 ( $33.14 \rightarrow 33.16$ ). This result, aligns with a growing body of recent work showing that the benefits of test-time compute are highly task-dependent (Liu et al., 2024b; Sprague et al., 2025; Gema et al., 2025) and does not necessarily improve role-playing ability (Feng et al., 2025). We hypothesize this is because CoT forces an overly rational deliberation that mismatches the often heuristic-driven nature of human responses in SimBench.

#### 4.3 RQ3: IMPACT OF TASK SELECTION ON SIMULATION FIDELITY

The 20 datasets in **SIMBENCH** correspond to very different tasks in terms of the aspects of human behavior that they measure (see §2.1). Breaking down simulation fidelity by dataset (Figure 3) reveals that **simulation fidelity varies substantially across tasks**. Models are most successful at simulating responses to standard survey questions regarding stated opinions, attitudes, and self-assessments (e.g., OpinionQA, Afrobarometer). However, performance degrades on tasks requiring the simulation of a *behavioral choice*, whether in economic gambles (Choices13k) or moral dilemmas (MoralMachine). This finding provides large-scale evidence for a “value-action gap” in LLMs, echoing recent work (Shen et al., 2025) which suggests a misalignment exists between LLM-generated value statements and their actions.

Finally, models struggle severely to capture human distributions on datasets involving traits or beliefs that conflict with standard alignment objectives. We observe extremely poor performance—often worse than a uniform baseline—on datasets measuring Machiavellianism (OSPsychMach), conspiratorial beliefs (ConspiracyCorr), or subjective humor (Jester). This aligns with a growing body of work (Liu et al., 2024a; Kumar et al., 2025; Yi et al., 2025) showing that alignment filters may inhibit the simulation of “atypical” or counter-normative human perspectives. While most LLMs exhibit these performance patterns, GPT-4.1 is a notable outlier, scoring exceptionally high (61.9) on OSPsychRWAS.

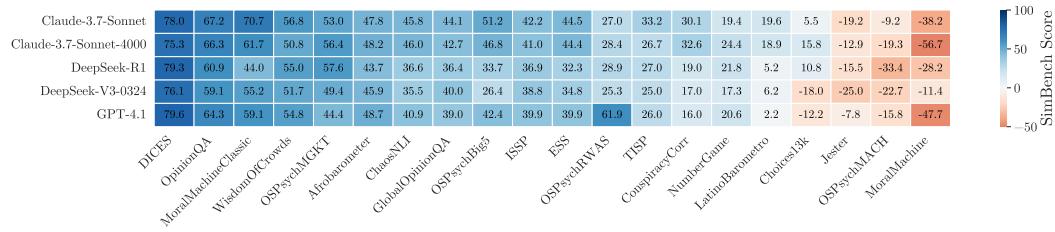


Figure 3: **Simulation fidelity by dataset** as measured by SIMBENCH score  $S$  for each of the 20 datasets in SimBenchPop. We show results for the top five models based on results in Table 1.

#### 4.4 RQ4: THE ALIGNMENT-SIMULATION TRADEOFF

Faithful simulation requires models to capture the full spectrum of human opinion, from strong consensus to widespread disagreement. We operationalize this “response plurality” using the normalized entropy of the human response distribution. Prior work suggests that standard alignment via instruction-tuning encourages confident, low-entropy outputs (Brown et al., 2020; Tian et al., 2023; Meister et al., 2025; Cruz et al., 2024), creating a potential conflict with simulating diverse human perspectives.

**An Empirical Tradeoff between Alignment and Plurality.** Our analysis provides strong evidence for this hypothesis, revealing a systematic tradeoff. As detailed in Appendix I, base models consistently outperform their instruction-tuned counterparts on high-entropy questions, while the inverse is true for low-entropy questions. To precisely quantify this effect, we compute the change in SIMBENCH score ( $\Delta S = S_{\text{instruct}} - S_{\text{base}}$ ) for 13 model pairs and plot this performance gain from post-training against human response entropy.

The result, shown in Figure 4, is a near-perfect negative linear relationship ( $r = -0.942$ ). This plot reveals two distinct regimes. On low-entropy questions where humans agree, post-training provides a

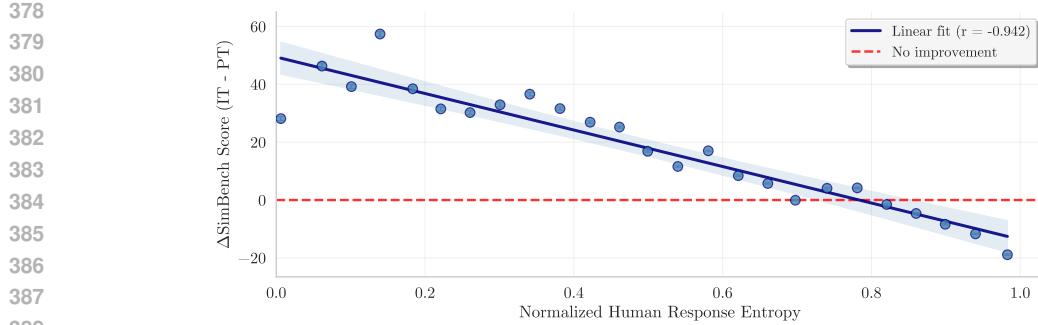


Figure 4: **The Alignment-Simulation Tradeoff: Instruction-Tuning Helps on Consensus Questions but Hurts on Diverse ones.** We bin questions by the normalized entropy of their human response distribution (x-axis) and plot the mean improvement in SIMBENCH Score when switching from a base model to its instruction-tuned counterpart (y-axis).

substantial benefit, improving the S-score by up to 40 points. This is alignment working as intended. However, as human disagreement (entropy) increases, the benefit of instruction-tuning systematically erodes, crossing a point of no-improvement around an entropy of 0.8. For questions with the highest plurality, post-training becomes actively detrimental, making the aligned model a *worse* simulator than its base counterpart.

This empirical finding is well-explained by the theoretical framework of reinforcement learning (RL) as Bayesian inference (Levine, 2018; Korbak et al., 2022). The pre-training objective of a base model typically minimizes a *mass-covering* KL divergence ( $D_{KL}(p\|q)$ ), which encourages the model ( $q$ ) to place probability mass wherever the true data distribution ( $p$ ) has mass. This process naturally leads to models that represent the full, multi-modal diversity of human language and opinion seen in their training data. In contrast, alignment via KL-regularized RL (e.g., RLHF) minimizes a *mode-seeking* KL divergence ( $D_{KL}(q\|\sigma)$ ). This objective incentivizes the model ( $q$ ) to find and concentrate its probability mass on a single, high-reward mode of the target preference distribution ( $\sigma$ ), even at the cost of ignoring other valid modes. Our results provide strong empirical validation of this theoretical distinction: alignment optimizes for a single “best” response, [effectively encouraging](#) the model to discard the pluralistic, high-entropy distributions characteristic of genuine human populations.

**Decomposing Instruction Tuning’s Dual Effects.** To formally test this mechanism, a causal mediation analysis (Appendix I.3) decomposes the effect of instruction-tuning into two larger, opposing forces: a large, positive *direct effect* on performance (**+6.46**), likely from improved instruction-following, and a significant, negative *indirect effect* (**-1.74**) mediated by the model’s reduced output entropy.

**Case Study: General-Purpose vs. Specialist Cognitive Tuning.** This formally-decomposed tradeoff is perfectly illustrated by comparing general-purpose instruction tuning with the specialist cognitive tuning of the Centaur models (Binz et al., 2025). Centaur models are Llama models fine-tuned on Psych-101, a large dataset of lab experiments, making our diverse SIMBENCH a powerful out-of-distribution test of their generalization. Both approaches improve simulation over the Llama-3.1-70B base model, but they do so via opposing mechanisms. **General-purpose instruction tuning** ( $S = 17.26$ ) leverages the helpful direct effect of alignment, excelling on low-entropy consensus tasks. In contrast, **specialist cognitive tuning** ( $S = 8.64$ ) improves performance by avoiding the harmful indirect effect, preserving the base model’s intrinsic ability to capture high-entropy pluralistic responses. The existence of these two distinct—and currently separate—paths to improving simulation underscores a key challenge and opportunity: the most faithful simulators of the future will likely need to synthesize the benefits of both general-purpose alignment and distribution-preserving cognitive modeling.

432 4.5 RQ5: SIMULATION ABILITY ACROSS PARTICIPANT GROUPS  
433  
434435 Many applications require simulating responses from specific demographic groups rather than general  
436 populations. Using SimBenchGrouped, we evaluate how LLM simulation ability changes when  
437 conditioned on specific demographic attributes.438 We measure this change as  $\Delta S = S_{grouped} - S_{ungrouped}$ , where  
439  $S_{ungrouped}$  is the SIMBENCH score for simulating the general  
440 population and  $S_{grouped}$  is the score when simulating a specific  
441 demographic group on the same question. A negative  $\Delta S$  indicates  
442 that the model’s simulation ability relative to the uniform  
443 baseline decreases when asked to simulate specific demographic  
444 groups.445 Importantly, for SimBenchGrouped, we specifically selected  
446 questions where human response distributions showed the highest  
447 variance across demographic groups (see §2.4). The observed  
448 degradation in simulation performance therefore likely  
449 represents an upper bound on the challenges LLMs face when  
450 simulating specific demographic groups. Our results in Table  
451 2 show that **LLMs struggle more with simulating specific**  
452 **demographic groups compared to general populations.** All  
453 evaluated models show negative mean  $\Delta S$  values, with degradation  
454 ranging from -1.27 for DeepSeek-V3-0324 to -4.61 for  
455 Claude-3.7-Sonnet-4000.456 The performance degradation varies substantially by demographic category. Models struggle most  
457 when simulating groups defined by religious attributes, with conditioning on “Religiosity/Practice”  
458 causing the largest decrease in simulation accuracy ( $\Delta S = -9.91$ ), followed by “Political Affiliation/Ideology”  
459 ( $\Delta S = -4.97$ ) and “Religion (Affiliation)” ( $\Delta S = -4.83$ ). In contrast, models  
460 maintain relatively better performance when simulating groups defined by “Gender” ( $\Delta S = -1.24$ )  
461 and “Age” ( $\Delta S = -1.50$ ).462 While these findings may not fully generalize to cases where demographic differences are less  
463 pronounced, they highlight potential limitations in how current LLMs capture the nuanced response  
464 patterns of specific demographic groups. We argue that such challenging benchmarks are crucial  
465 for identifying areas where improvements are most needed, particularly for applications that aim to  
466 model the behaviors of specific subpopulations.470 4.6 RQ6: SIMULATION ABILITY VS. GENERAL CAPABILITIES  
471  
472473 Finally, we analyze the relationship between LLM simulation ability and more general model  
474 capabilities by correlating performance on SIMBENCH with popular LLM capability benchmarks.  
475 We collect performance data for eight models on five benchmarks representing distinct capabilities  
476 and calculate the Pearson correlation with their SIMBENCH scores (see Appendix J for detailed  
477 implementation details and scatter plots).478 We find that SIMBENCH performance correlates most strongly with benchmarks requiring deep and  
479 broad knowledge-intensive reasoning, such as **MMLU-Pro** ( $r = 0.94$ ) and **GPQA Diamond** ( $r = 0.86$ ).  
480 This relationship is weaker for general chatbot helpfulness as measured by **Chatbot Arena ELO** ( $r =$   
481 **0.71**) and instruction following (**IF-Eval**,  $r = 0.79$ ). Crucially, the correlation is substantially weaker  
482 for narrow, specialized skills like advanced mathematics (**OTIS AIME**,  $r = 0.48$ ). We posit that  
483 accurately simulating human behavior is a complex capability rooted in broad, knowledge-intensive  
484 reasoning, which aligns with the diverse social and behavioral topics in SIMBENCH. The weaker  
485 correlations with Chatbot Arena and OTIS AIME show that neither general conversational ability  
nor narrow problem-solving skills are sufficient proxies for strong simulation performance.435 Table 2: **Ungrouped vs. grouped**  
436 **simulation performance ( $\Delta S$ ).**

Category	$\Delta S$
<b>By Models</b>	
Claude-3.7-Sonnet	-3.13
Claude-3.7-Sonnet-4000	-4.61
DeepSeek-R1	-3.79
DeepSeek-V3-0324	-1.27
GPT-4.1	-3.94
<b>By Demographics</b>	
Religiosity/Practice	-9.91
Political Affil./Ideology	-4.97
Religion (Affiliation)	-4.83
Income/Social Standing	-4.51
Domicile/Urbanicity	-3.17
Employment Status	-3.03
Education	-2.55
Marital Status	-1.80
Age	-1.50
Gender	-1.24

## 486 5 RELATED WORK

488 **Human Behavior Simulation with LLMs** LLMs as human behavior simulators have attracted  
 489 significant interdisciplinary attention. Researchers have evaluated their efficacy across political  
 490 science (Argyle et al., 2023; Bisbee et al., 2024; Dominguez-Olmedo et al., 2024), psychology (Aher  
 491 et al., 2023; Binz et al., 2025; Manning et al., 2024; Hewitt et al., 2024), economics (Horton, 2023;  
 492 Aher et al., 2023), and computer science applications (Hu & Collier, 2024; Dong et al., 2024; Hu &  
 493 Collier, 2025; Park et al., 2023). Evidence regarding LLMs’ simulation fidelity remains mixed, with  
 494 some studies reporting promising results (Argyle et al., 2023) while others identify critical limitations,  
 495 including homogenized group representations (Cheng et al., 2023; Wang et al., 2025), **a tendency**  
 496 **toward hyper-rationality rather than human-like error (Liu et al., 2025)**, and deterministic rather than  
 497 distributional predictions (Park et al., 2024b).

498 Existing work has predominantly focused on individual-level simulation with minimal demographic  
 499 conditioning, typically evaluating only one or two models in narrowly defined contexts. SIMBENCH  
 500 addresses these limitations by providing a comprehensive benchmark for group-level simulation  
 501 across diverse domains with systematic demographic conditioning and standardized metrics. The  
 502 benchmark’s distributional evaluation framework (using Total Variation distance) captures how  
 503 accurately models represent the full spectrum of human response variation – an approach advocated  
 504 by researchers in both simulation (Anthis et al., 2025) and general LLM evaluation (Ying et al.,  
 505 2025). For broader context on this emerging field, we refer readers to recent comprehensive surveys  
 506 (Kozlowski & Evans, 2024; Olteanu et al., 2025; Anthis et al., 2025).

507 Appendix M continues our discussion of related work.

## 509 6 CONCLUSION AND FUTURE WORK

511 For LLM simulations to become reliable tools for the social and behavioral sciences, their fidelity  
 512 to real human behavior must be measurable. However, prior evaluations have been fragmented,  
 513 hindering systematic progress. To address this, we introduce SIMBENCH, the first large-scale,  
 514 standardized benchmark for group-level human behavior simulation. By unifying 20 diverse datasets,  
 515 SIMBENCH provides the necessary infrastructure to robustly evaluate and compare LLM simulators.

516 Using this benchmark, we provide the first systematic analysis of this capability, showing that  
 517 even SOTA LLMs have limited simulation ability, performance scales log-linearly with model size,  
 518 and there exists a fundamental tradeoff between standard alignment and simulating diverse human  
 519 opinions. We further reveal that models struggle more when simulating specific demographic groups.  
 520 We also show that strong simulation ability correlates with deep, knowledge-intensive reasoning.  
 521 While significant progress is needed, SIMBENCH makes this progress measurable, providing an open  
 522 foundation to accelerate the development of more faithful LLM simulators.

523 Future research should build on this foundation by expanding beyond standardized formats to evaluate  
 524 **interactive and open-ended behavioral simulations**. Additionally, benchmarking generative  
 525 simulators against **supervised machine learning baselines** would help clarify the comparative  
 526 value of LLM-based simulation. Finally, addressing the observed tradeoff between alignment and  
 527 simulation fidelity requires developing **distribution-preserving alignment** techniques, alongside  
 528 further investigation into the **causal mechanisms** linking other model capabilities to simulation  
 529 performance.

530  
 531  
 532  
 533  
 534  
 535  
 536  
 537  
 538  
 539

540 7 ETHICS STATEMENT  
541542 SimBench’s primary purpose is to benchmark LLMs’ ability to simulate human behavior. While  
543 advancements in LLM simulation capabilities can support helpful applications such as pre-testing  
544 policies, these do not come without risks of misrepresentation and dual use.  
545546 7.1 RESPONSIBLE USE AND ACKNOWLEDGMENT OF LIMITATIONS  
547548 First and foremost, due to the observed limited simulation ability of state-of-the-art LLMs, we  
549 caution against relying on LLM-powered simulations of human behavior for tasks where downstream  
550 harm is possible. Even as models improve, substituting algorithmic approximations for authentic  
551 human participation carries the risk of disadvantaging under-represented / marginalized communities  
552 by removing their opportunities to directly shape decisions that affect them. Furthermore, while  
553 benchmarks like **SIMBENCH** help measure simulation capabilities, we must be careful not to mistake  
554 increasing benchmark performance for genuine understanding of complex human behavior.  
555556 7.2 DATA PROVENANCE AND TRANSFORMATIVE USE  
557558 The creation of **SIMBENCH** from 20 diverse sources was guided by a commitment to responsible  
559 data handling. Our curation process prioritized datasets with clear and permissive terms. As a result,  
560 17 out of the 20 datasets are governed by explicit permissive licenses (e.g., Creative Commons, MIT).  
561 For the few remaining datasets that are publicly available for research without an explicit license, we  
562 apply a consistent framework built on the principle of transformative use.  
563564 1. **Transformative Use.** **SIMBENCH** does not contain or redistribute any raw, individual-level  
565 participant data. It is a new, derivative work consisting of aggregated, non-reversible group-level  
566 distributions. This process protects the privacy of the original human subjects.  
567 2. **Multi-Level Licensing.** Our public release includes a detailed `LICENSE` file. **The SIMBENCH**  
568 **framework** (our code and pipeline) is permissively licensed (e.g., CC-BY-NC-SA 4.0). For each  
569 of the **20 constituent datasets**, the documentation explicitly lists the original source, its specific  
570 license or terms of use, and a clear statement clarifying its status as an aggregated, derivative work  
571 whose original terms should still be consulted.  
572573 7.3 SCOPE OF REPRESENTATION AND INTERSECTIONAL ANALYSIS  
574575 While **SIMBENCH** includes diverse demographic groups, it can not adequately support simulations of  
576 intersectional identities due to sample size limitations. By conditioning on one demographic variable  
577 at a time, we cannot systematically assess how well models handle the rich overlap of identities  
578 (e.g., “older Latinx women,” “young Black men”). This was a deliberate methodological choice  
579 to maintain the statistical integrity of the ground-truth distributions, as small intersectional group  
580 sizes make it difficult to combine multiple characteristics simultaneously due to increasing sampling  
581 noise in response distributions. Yet intersectional simulation is precisely where societal biases and  
582 model limitations often emerge, making this an important direction for future work. Additionally, the  
583 conditional prompting approach we use conceptualizes simplistic human populations and may thus  
584 fail to appropriately account for nuances of individual behavior.  
585586 7.4 CONCLUSION  
587588 Nevertheless, we believe **SIMBENCH** is an important step toward making LLM simulation progress  
589 measurable and raising awareness of state-of-the-art model blind spots. Together, we hope this will  
590 ultimately create accountability for models deployed in socially sensitive contexts.  
591592 8 REPRODUCIBILITY STATEMENT  
593594 To ensure the reproducibility of our findings and to facilitate future research, we make all components  
595 of our work publicly available. The complete **SIMBENCH** benchmark, including all 20 processed  
596 datasets, our curated SimBenchPop and SimBenchGrouped splits, and detailed data cards, will  
597 be released on the Hugging Face Hub. Our codebase will be available in an open-source GitHub  
598 repository.  
599

594 repository. We provide detailed descriptions of our experimental setup, including the exact prompts  
595 used for both base and instruction-tuned models, in Appendix D. We further provide an empirical  
596 validation of our elicitation methodology choice in Appendix E.  
597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648 REFERENCES  
649

650 Afrobarometer. Afrobarometer data, all countries (39), round 9, 2023. <http://www.afrobarometer.org>, 2023. Accessed: March 2025.

651

652 Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate  
653 multiple humans and replicate human subject studies. In Andreas Krause, Emma Brunskill,  
654 Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of  
655 the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine  
656 Learning Research*, pp. 337–371. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/aher23a.html>.

657

658 Georg Ahnert, Max Pellert, David Garcia, and Markus Strohmaier. Extracting affect aggregates from  
659 longitudinal social media data with temporal adapters for large language models. *arXiv preprint  
660 arXiv:2409.17990*, 2024.

661

662 Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, James Evans,  
663 Erik Brynjolfsson, and Michael Bernstein. Llm social simulations are a promising research method.  
664 *arXiv preprint arXiv:2504.02234*, 2025.

665

666 Anthropic. Claude 3.7 sonnet and claude code, February 2025. URL <https://www.anthropic.com/news/claude-3-7-sonnet>. Accessed: 2025-05-14.

667

668

669 Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David  
670 Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*,  
671 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

672

673 Lora Aroyo, Alex S. Taylor, Mark Díaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-García,  
674 Vinodkumar Prabhakaran, and Ding Wang. Dices dataset: diversity in conversational ai evaluation  
675 for safety. In *Proceedings of the 37th International Conference on Neural Information Processing  
676 Systems*, NIPS ’23, Red Hook, NY, USA, 2023. Curran Associates Inc.

677

678 Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-  
679 François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64,  
680 2018.

681

682 Edmond Awad, Sohan Dsouza, Azim Shariff, Iyad Rahwan, and Jean-François Bonnefon. Universals  
683 and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the  
National Academy of Sciences*, 117(5):2332–2337, 2020.

684

685 Eric Bigelow and Steven T Piantadosi. A large dataset of generalization patterns in the number game.  
686 *Journal of Open Psychology Data*, 4(1):e4–e4, 2016.

687

688 Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno,  
689 Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. A foundation model to predict  
690 and capture human cognition. *Nature*, pp. 1–8, 2025.

691

692 James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic  
693 replacements for human survey data? the perils of large language models. *Political Analysis*, 32  
(4):401–416, 2024. doi: 10.1017/pan.2024.5.

694

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

698

699 Myra Cheng, Tiziano Piccardi, and Diyi Yang. CoMPosT: Characterizing and evaluating caricature in  
700 LLM simulations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023  
701 Conference on Empirical Methods in Natural Language Processing*, pp. 10853–10875, Singapore,  
702 December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.  
699. URL <https://aclanthology.org/2023.emnlp-main.669/>.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

André F. Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 97378–97407. Curran Associates, Inc., 2024. URL [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/b0a4b3e384b4554e65a47ad1f6b0310a-Paper-Datasets\\_and\\_Benchmarks\\_Track.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/b0a4b3e384b4554e65a47ad1f6b0310a-Paper-Datasets_and_Benchmarks_Track.pdf).

DeepSeek-AI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 45850–45878. Curran Associates, Inc., 2024. URL [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/515c62809e0a29729d7eec26e2916fc0-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/515c62809e0a29729d7eec26e2916fc0-Paper-Conference.pdf).

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. Can LLM be a personalized judge? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10126–10141, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.592. URL <https://aclanthology.org/2024.findings-emnlp.592/>.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=z116jLb91v>.

Adam Enders, Casey Klofstad, Amanda Diekman, Hugo Drochon, Joel Rogers de Waal, Shane Littrell, Kamal Premaratne, Daniel Verdear, Stefan Wuchty, and Joseph Uscinski. The sociodemographic correlates of conspiracism. *Scientific reports*, 14(1):14184, 2024.

European Social Survey European Research Infrastructure (ESS ERIC). ESS11 - Integrated File, Edition 2.0 [Data set], 2024. URL [https://doi.org/10.21338/ess11e02\\_0](https://doi.org/10.21338/ess11e02_0).

Xiachong Feng, Longxu Dou, and Lingpeng Kong. Reasoning does not necessarily improve role-playing ability. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10301–10314, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.537. URL <https://aclanthology.org/2025.findings-acl.537/>.

Clémantine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open llm leaderboard v2. [https://huggingface.co/spaces/open-llm-leaderboard/open\\_llm\\_leaderboard](https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard), 2024.

Aryo Pradipta Gema, Alexander Hägele, Runjin Chen, Andy Ardit, Jacob Goldman-Wetzler, Kit Fraser-Taliente, Henry Sleight, Linda Petrini, Julian Michael, Beatrice Alex, et al. Inverse scaling in test-time compute. *arXiv preprint arXiv:2507.14417*, 2025.

Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *information retrieval*, 4:133–151, 2001.

756 Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood,  
 757 Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam  
 758 Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason  
 759 Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan Choi, Kevin Tobia,  
 760 Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger,  
 761 Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams,  
 762 Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. Legalbench: A collaboratively built  
 763 benchmark for measuring legal reasoning in large language models. In *Advances in Neu-*  
 764 *ral Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*,  
 765 2023. URL [https://proceedings.neurips.cc/paper\\_files/paper/2023/hash/89e44582fd28ddfea1ea4dc0ebbf4b0-Abstract-Datasets\\_and\\_Benchmarks.html](https://proceedings.neurips.cc/paper_files/paper/2023/hash/89e44582fd28ddfea1ea4dc0ebbf4b0-Abstract-Datasets_and_Benchmarks.html).

766

767 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu  
 768 Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou,  
 769 Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei  
 770 Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie  
 771 Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li,  
 772 H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li,  
 773 Jingchang Chen, Jingyang Yuan, Jiniao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang,  
 774 Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean  
 775 Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan  
 776 Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming  
 777 Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi  
 778 Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu,  
 779 Shangyan Zhou, Shanhua Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou,  
 780 Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang,  
 781 Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao,  
 782 Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin  
 783 Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin,  
 784 Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu  
 785 Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao  
 786 Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi  
 787 Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan  
 788 Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu,  
 789 Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying  
 790 Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie,  
 791 Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun  
 792 Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang,  
 793 and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning.  
 794 *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL  
<https://doi.org/10.1038/s41586-025-09422-z>.

795 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob  
 796 Steinhardt. Measuring massive multitask language understanding. In *International Confer-*  
 797 *ence on Learning Representations*, 2021. URL <https://openreview.net/forum?id=d7KBjmI3GmQ>.

798

799 Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. Predicting re-  
 800 sults of social science experiments using large language models, August 2024.  
 801 URL <https://samim.io/dl/Predicting%20results%20of%20social%20science%20experiments%20using%20large%20language%20models.pdf>.

802

803 John J Horton. Large language models as simulated economic agents: What can we learn from homo  
 804 silicus? Technical report, National Bureau of Economic Research, 2023.

805

806 Tiancheng Hu and Nigel Collier. Quantifying the persona effect in LLM simulations. In Lun-  
 807 Wei Ku, Andre Martins, and Vivek Srikanth (eds.), *Proceedings of the 62nd Annual Meeting*  
 808 *of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10289–10307,  
 809 Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.554. URL <https://aclanthology.org/2024.acl-long.554/>.

810 Tiancheng Hu and Nigel Collier. inews: A multimodal dataset for modeling personalized affective  
 811 responses to news. *arXiv preprint arXiv:2503.03335*, 2025.

812

813 ISSP Research Group. International social survey programme: Social networks and social  
 814 resources - issp 2017. GESIS Data Archive, Cologne. ZA6980 Data file Version 2.0.0,  
 815 <https://doi.org/10.4232/1.13322>, 2019.

816 ISSP Research Group. International social survey programme: Religion iv - issp 2018. GESIS Data  
 817 Archive, Cologne. ZA7570 Data file Version 2.1.0, <https://doi.org/10.4232/1.13629>, 2020.

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

820

821 ISSP Research Group. International social survey programme: Environment iv - issp 2020. GESIS,  
 822 Cologne. ZA7650 Data file Version 2.0.0, <https://doi.org/10.4232/1.14153>, 2023.

823 ISSP Research Group. Za8000 international social survey programme: Health and health care ii - issp  
 824 2021. GESIS, Cologne. ZA8000 Data file Version 2.0.0, <https://doi.org/10.4232/5.ZA8000.2.0.0>,  
 825 2024.

826

827 Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language  
 828 models know? on the calibration of language models for question answering. *Transactions of the*  
 829 *Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl\_a\_00407. URL  
 830 <https://aclanthology.org/2021.tacl-1.57/>.

831 Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. In *Pro-*  
 832 *ceedings of the 56th Annual ACM Symposium on Theory of Computing*, STOC 2024, pp. 160–171,  
 833 New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703836. doi:  
 834 10.1145/3618260.3649777. URL <https://doi.org/10.1145/3618260.3649777>.

835 Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and  
 836 Andrew Wilson. Calibration-tuning: Teaching large language models to know what they don't  
 837 know. In Raúl Vázquez, Hande Celikkanat, Dennis Ulmer, Jörg Tiedemann, Swabha Swayamdipta,  
 838 Wilker Aziz, Barbara Plank, Joris Baan, and Marie-Catherine de Marneffe (eds.), *Proceedings of*  
 839 *the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pp. 1–14, St Julians, Malta,  
 840 March 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.uncertainlp-1.1/>.

841

842 Tomasz Korbak, Ethan Perez, and Christopher Buckley. RL with KL penalties is better viewed as  
 843 Bayesian inference. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the*  
 844 *Association for Computational Linguistics: EMNLP 2022*, pp. 1083–1091, Abu Dhabi, United  
 845 Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.77. URL <https://aclanthology.org/2022.findings-emnlp.77/>.

846

847 Austin C. Kozlowski and James Evans. Simulating subjects: The promise and peril of ai stand-ins for  
 848 social agents and interactions, September 2024. URL [https://www.researchgate.net/publication/383972192\\_Simulating\\_Subjects\\_The\\_Promise\\_and\\_Peril\\_of\\_AI\\_Stand-ins\\_for\\_Social\\_Agents\\_and\\_Interactions](https://www.researchgate.net/publication/383972192_Simulating_Subjects_The_Promise_and_Peril_of_AI_Stand-ins_for_Social_Agents_and_Interactions). Preprint.

849

850

851

852 Sai Adith Senthil Kumar, Hao Yan, Saipavan Perepa, Murong Yue, and Ziyu Yao. Can llms simulate  
 853 personas with reversed performance? a benchmark for counterfactual instruction following, 2025.  
 854 URL <https://arxiv.org/abs/2504.06460>.

855

856 Latinobarómetro. Latinobarómetro 2023. <http://www.latinobarometro.org>, 2023. Ac-  
 857 cessed: March 2025.

858

859 Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi,  
 860 Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kociský, Sebastian Ruder, Dani Yogatama,  
 861 Kris Cao, Susannah Young, and Phil Blunsom. Mind the gap: Assessing temporal generalization in  
 862 neural language models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman  
 863 Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29348–29363.  
 864 Curran Associates, Inc., 2021. URL [https://proceedings.neurips.cc/paper\\_files/paper/2021/file/f5bf0ba0a17ef18f9607774722f5698c-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/f5bf0ba0a17ef18f9607774722f5698c-Paper.pdf).

864 Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review.  
 865 *arXiv preprint arXiv:1805.00909*, 2018.  
 866

867 Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human  
 868 falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of*  
 869 *the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,  
 870 pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi:  
 871 10.18653/v1/2022.acl-long.229. URL [https://aclanthology.org/2022.acl-long.](https://aclanthology.org/2022.acl-long.229/)  
 872 229/.

873 Andy Liu, Mona Diab, and Daniel Fried. Evaluating large language model biases in persona-  
 874 steered generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the*  
 875 *Association for Computational Linguistics: ACL 2024*, pp. 9832–9850, Bangkok, Thailand, August  
 876 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.586. URL  
 877 <https://aclanthology.org/2024.findings-acl.586/>.

878 Ryan Liu, Jiayi Geng, Addison J Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L Griffiths.  
 879 Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes  
 880 humans worse. *arXiv preprint arXiv:2410.21333*, 2024b.

881

882 Ryan Liu, Jiayi Geng, Joshua Peterson, Ilia Sucholutsky, and Thomas L. Griffiths. Large language  
 883 models assume people are more rational than we really are. In *The Thirteenth International*  
 884 *Conference on Learning Representations*, 2025. URL [https://openreview.net/forum?](https://openreview.net/forum?id=dAeET8gxqg)  
 885 [id=dAeET8gxqg](https://openreview.net/forum?id=dAeET8gxqg).

886 Benjamin S Manning, Kehang Zhu, and John J Horton. Automated social science: Language models  
 887 as scientist and subjects. Technical report, National Bureau of Economic Research, 2024.

888

889 Niels G Mede, Viktoria Cologna, Sebastian Berger, John Besley, Cameron Brick, Marina Joubert,  
 890 Edward W Maibach, Sabina Mihelj, Naomi Oreskes, Mike S Schäfer, et al. Perceptions of science,  
 891 science communication, and climate change attitudes in 68 countries—the tisp dataset. *Scientific*  
 892 *data*, 12(1):114, 2025.

893

894 Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. Benchmarking distributional alignment  
 895 of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the*  
 896 *2025 Conference of the Nations of the Americas Chapter of the Association for Computational*  
 897 *Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 24–49, Albuquerque,  
 898 New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6.  
 899 URL <https://aclanthology.org/2025.naacl-long.2/>.

900

901 Meta AI. Introducing Llama 3.1: Our most capable models to date, 2024. URL <https://ai.meta.com/blog/meta-llama-3-1/>. Accessed: 2025-05-14.

902

903 Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions  
 904 on natural language inference data? In Bonnie Webber, Trevor Cohn, Yulan He, and Yang  
 905 Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*  
 906 *Processing (EMNLP)*, pp. 9131–9143, Online, November 2020. Association for Computational  
 907 Linguistics. doi: 10.18653/v1/2020.emnlp-main.734. URL <https://aclanthology.org/2020.emnlp-main.734/>.

908

909 Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities  
 910 of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

911

912 Alexandra Olteanu, Solon Barocas, Su Lin Blodgett, Lisa Egede, Alicia DeVrio, and Myra Cheng.  
 913 Ai automatons: Ai systems intended to imitate humans. *arXiv preprint arXiv:2503.02250*, 2025.

914

915 OpenAI. Introducing GPT-4.1 in the API, 2025a. URL <https://openai.com/index/gpt-4-1/>. Accessed: 2025-05-14.

916

917 OpenAI. Introducing openai o3 and o4-mini, April 2025b. URL <https://openai.com/index/introducing-o3-and-o4-mini/>. Accessed: 2025-05-14.

918 Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S.  
919 Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the*  
920 *36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New  
921 York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi:  
922 10.1145/3586183.3606763. URL <https://doi.org/10.1145/3586183.3606763>.

923 Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel  
924 Morris, Robb Willer, Percy Liang, and Michael S Bernstein. Generative agent simulations of 1,000  
925 people. *arXiv preprint arXiv:2411.10109*, 2024a.

927 Peter S. Park, Philipp Schoenegger, and Chongyang Zhu. Diminished diversity-of-thought in a  
928 standard large language model. *Behavior Research Methods*, 56:5754–5770, 2024b. doi: 10.  
929 3758/s13428-023-02307-x. URL <https://link.springer.com/article/10.3758/s13428-023-02307-x>.

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

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

938 Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python.  
939 In Stéfan van der Walt and Jarrod Millman (eds.), *Proceedings of the 9th Python in Science*  
940 Conference, pp. 92–96, 2010. doi: 10.25080/Majora-92bf1922-011.

941 Hua Shen, Nicholas Clark, and Tanu Mitra. Mind the value-action gap: Do LLMs act in alignment  
942 with their values? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet  
943 Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language*  
944 *Processing*, pp. 3097–3118, Suzhou, China, November 2025. Association for Computational  
945 Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.154. URL <https://aclanthology.org/2025.emnlp-main.154/>.

947 Camelia Simoiu, Chiraag Sumanth, Alok Mysore, and Sharad Goel. Studying the “wisdom of crowds”  
948 at scale. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*,  
949 volume 7, pp. 171–179, 2019.

951 Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann  
952 Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-  
953 thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference*  
954 *on Learning Representations*, 2025. URL <https://openreview.net/forum?id=w6n1cS8Kkn>.

956 Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej,  
957 Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical  
958 report. *arXiv preprint arXiv:2503.19786*, 2025.

959 Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn,  
960 and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence  
961 scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino,  
962 and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural*  
963 *Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational  
964 Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL <https://aclanthology.org/2023.emnlp-main.330/>.

966 Linda Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. Do llms  
967 exhibit human-like response biases? a case study in survey design. *Transactions of the Association*  
968 *for Computational Linguistics*, 12:1011–1026, 09 2024. ISSN 2307-387X. doi: 10.1162/tacl\_a\_00685. URL [https://doi.org/10.1162/tacl\\_a\\_00685](https://doi.org/10.1162/tacl_a_00685).

971 Vals AI, Inc. Mmlu pro benchmark. [https://www.vals.ai/benchmarks/mmlu\\_pro-04-15-2025](https://www.vals.ai/benchmarks/mmlu_pro-04-15-2025), April 2025. Last updated 15 April 2025.

972 Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE:  
 973 A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen,  
 974 Grzegorz Chrupała, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop Black-*  
 975 *boxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium,  
 976 November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL  
 977 <https://aclanthology.org/W18-5446/>.

978 Angelina Wang, Jamie Morgenstern, and John P Dickerson. Large language models that replace hu-  
 979 man participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*,  
 980 pp. 1–12, 2025.

981 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi,  
 982 Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language  
 983 models. In *Proceedings of the 36th International Conference on Neural Information Processing*  
 984 *Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

985 Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi,  
 986 Pierrick Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von  
 987 Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama  
 988 Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language  
 989 processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on*  
 990 *Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online,  
 991 October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6.  
 992 URL <https://aclanthology.org/2020.emnlp-demos.6/>.

993 An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li,  
 994 Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint*  
 995 *arXiv:2412.15115*, 2024.

996 Zihao Yi, Qingxuan Jiang, Ruotian Ma, Xingyu Chen, Qu Yang, Mengru Wang, Fanghua Ye, Ying  
 997 Shen, Zhaopeng Tu, Xiaolong Li, and Linus. Too good to be bad: On the failure of llms to role-play  
 998 villains, 2025. URL <https://arxiv.org/abs/2511.04962>.

999 Lance Ying, Katherine M Collins, Lionel Wong, Ilia Sucholutsky, Ryan Liu, Adrian Weller, Tianmin  
 1000 Shu, Thomas L Griffiths, and Joshua B Tenenbaum. On benchmarking human-like intelligence in  
 1001 machines. *arXiv preprint arXiv:2502.20502*, 2025.

1002 Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving  
 1003 few-shot performance of language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of*  
 1004 *the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine*  
 1005 *Learning Research*, pp. 12697–12706. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/zhao21c.html>.

1006 Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. On the calibration  
 1007 of large language models and alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.),  
 1008 *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9778–9795, Sin-  
 1009 gapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.  
 1010 findings-emnlp.654. URL <https://aclanthology.org/2023.findings-emnlp.654/>.

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

1026 **A LIMITATIONS**

1028 **The Scope of Benchmarking Human Behavior** The challenge of benchmarking human behavior  
 1029 lies in its vastness. No single study, nor dozens, could ever capture its full complexity. We operational-  
 1030 ize this challenge by focusing on a core set of fundamental cognitive and social tasks widely studied  
 1031 in the behavioral sciences: decision-making, self-assessment, judgment, and problem-solving. This  
 1032 focus necessarily means that SIMBENCH does not capture the full complexity of human interaction,  
 1033 such as embodied or multi-turn social dynamics. As the first large-scale benchmark for group-level  
 1034 simulation, SIMBENCH provides the essential infrastructure to establish robust baselines, uncover  
 1035 fundamental properties like [scaling trends](#), and map the frontier for future work on more complex,  
 1036 interactive simulations.

1037 **Scope of Representativeness** Although SIMBENCH spans 20 diverse datasets, the combined  
 1038 sample does (and can) not fully represent any single population in its full complexity. Many  
 1039 geographic regions are still underrepresented or entirely absent, potentially limiting generalizability to  
 1040 populations with different cultural backgrounds and preferences. Even within countries, demographic  
 1041 representativeness may vary, as only a subset of our 20 datasets are based on nationally representative  
 1042 sampling techniques. Each dataset carries its own statistical uncertainty. Opt-in samples and  
 1043 crowdsourced data (e.g., from Amazon Mechanical Turk) may have larger margins of error than  
 1044 nationally representative surveys, potentially affecting the benchmark’s precision for certain questions.  
 1045 We view these limitations as opportunities for collaborative extension of SIMBENCH to improve  
 1046 global coverage and representativeness over time.

1047 **Temporal Dimensions** The current version of SIMBENCH utilizes static datasets that capture  
 1048 human behavior at specific points in time. This approach allows for systematic evaluation across  
 1049 domains but cannot yet assess how well LLMs simulate evolving preferences, opinion shifts, or  
 1050 behavioral adaptation – all fundamental aspects of human behavior. Future iterations of SIMBENCH  
 1051 could incorporate longitudinal data to address these dynamic aspects of human behavior and expand  
 1052 the benchmark’s evaluative capacity.

1053 **Task Format Considerations** SIMBENCH currently focuses on multiple-choice, single-answer,  
 1054 single-turn questions and interactions. This standardized format enables systematic comparison  
 1055 across diverse domains but necessarily excludes more complex behavioral simulations including  
 1056 multi-step decision processes and interactive social dynamics. We see this as a pragmatic starting  
 1057 point that establishes foundational evaluation capabilities while inviting future extensions to capture  
 1058 more nuanced aspects of human behavior.

1059 **Training Data Overlap** A fundamental challenge for all LLM evaluation is the potential for  
 1060 training data contamination. While this cannot be definitively ruled out without full access to model  
 1061 training corpora, several aspects of our methodology and empirical findings suggest this risk is  
 1062 substantially mitigated for SIMBENCH:

- 1063 1) Much of our source data is not easily ingested by standard web scrapes. Virtually all of our source  
 1064 data are primarily distributed as structured data files in specialized formats like R or SAS in academic  
 1065 archives. This makes data contamination much less likely than for generic web text, as these files  
 1066 cannot be meaningfully read or interpreted as plain text by standard scraping tools.
- 1067 2) SimBench’s core task is not to recall a fact but to predict a response distribution for a specific  
 1068 demographic subgroup (e.g., “women in Slovakia”). Even if a model’s training data included  
 1069 thousands of individual survey responses, it would still need to learn, without supervision, how to  
 1070 aggregate these individual points into a coherent distribution for an arbitrary, specified subgroup.  
 1071 This is a sophisticated, zero-shot social reasoning skill that is unlikely to emerge from simply seeing  
 1072 the raw data.
- 1073 3) On datasets that are most likely to appear in training data (e.g., US-centric OpinionQA), even the  
 1074 best-performing models achieve an S-score of only 60, far from the 100-point maximum. If models  
 1075 had memorized this benchmark, we would expect scores far closer to perfect. This clear performance  
 1076 ceiling demonstrates that our benchmark is testing a genuine capability rather than memorization.

1080 4) The consistent scaling patterns we observe across diverse datasets suggest genuine simulation  
 1081 capabilities rather than artifacts of training data overlap.

1082 Nevertheless, we acknowledge that data contamination remains a fundamental challenge in LLM  
 1083 evaluation, and future work should develop more robust methods to detect and quantify its impact.  
 1084 We include this consideration for completeness while believing it unlikely to significantly impact our  
 1085 current findings.

## 1087 B DATASET CURATION AND EXCLUSION RATIONALE

1090 As described in §2.1, our dataset curation process involved a systematic review of numerous prominent  
 1091 datasets in the social and behavioral sciences. While our search was extensive, many promising  
 1092 candidates were ultimately excluded for failing to meet our strict inclusion criteria for a redistributable,  
 1093 multiple-choice benchmark.

1094 Table 3 provides an illustrative list of well-known datasets that were considered during this review  
 1095 but not included in the final SIMBENCH collection. This list is not exhaustive but serves to highlight  
 1096 the common methodological and logistical challenges that arise when creating a benchmark from  
 1097 existing scientific data, such as restrictive licensing, complex experimental designs, and non-standard  
 1098 response formats. By documenting these exclusion rationales, we aim to provide transparency into  
 1099 our curation process and offer a resource for future benchmark creators.

1100 1101 Table 3: Examples of Datasets Considered and Excluded from SIMBENCH.

1102 <b>Dataset</b>	1103 <b>Exclusion Reason</b>
1104 Understanding America Study	<i>Licensing Restrictions</i>
1105 ManyLabs & ManyLabs 2	<i>Complex treatment condition</i>
1106 General Social Survey	<i>Licensing Restrictions</i>
1107 Demographic and Health Surveys	<i>Licensing Restrictions</i>
1108 Global Preferences Survey	<i>Licensing Restrictions</i>
1109 World Risk Poll	<i>No raw question wording available</i>
1110 Yourmorals.org	<i>Licensing Restrictions</i>
1111 Asian Barometer	<i>Licensing Restrictions</i>
1112 The Glasgow Norms	<i>Licensing Restrictions</i>
1113 MovieLens	<i>Licensing Restrictions</i>
1114 Health Information National Trends Survey	<i>Licensing Restrictions</i>
1115 BBC Big Personality Test	<i>Licensing Restrictions</i>
1116 Time-sharing Experiments for the Social Sciences	<i>Often complex treatment conditions</i>
1117 Project Implicit	<i>Hard to model reaction time in LLMs</i>
1118 Children’s Worlds Survey	<i>Licensing Restrictions</i>
1119 Monitoring the Future Survey	<i>Licensing Restrictions</i>
1120 TIMSS	<i>Individual test items are not accessible</i>
1121 UMD-OurDataHelps	<i>Free text response</i>
1122 MobLab dataset	<i>Too few questions; lack detailed demographics data</i>

## 1126 C SIMBENCHPOP AND SIMBENCHGROUPED SAMPLING DETAILS

1128 We curated data at two levels of grouping granularity, corresponding to our two main benchmark  
 1129 splits: **SimBenchPop** and **SimBenchGrouped**.

1131 **SimBenchPop** measures LLMs’ ability to simulate responses of broad, diverse human populations.  
 1132 We include all questions from all 20 datasets in SimBench, combining each question with its dataset-  
 1133 specific default grouping prompt (e.g., “You are an Amazon Mechanical Turk worker based in the  
 United States”). We sample up to 500 questions per dataset to ensure representativeness while

1134 keeping the benchmark manageable. For each test case, we aggregate individual responses across all  
 1135 participants in the dataset to create population-level response distributions. This approach creates a  
 1136 benchmark that represents population-level responses across diverse domains while maintaining a  
 1137 reasonable size of 7,167 test cases.

1138 For **SimBenchGrouped**, we focus only on five large-scale survey datasets with rich demographic in-  
 1139 formation and sufficient sample sizes: OpinionQA, ESS, Afrobarometer, ISSP, and LatinoBarometro.  
 1140 Our sampling approach prioritizes questions showing meaningful demographic variation. For each  
 1141 dataset, we identify available grouping variables (e.g., age, gender, country) with sufficient group  
 1142 sizes to form meaningful response distributions. We calculate the variance of responses across  
 1143 demographic groups for each question and rank questions by their variance scores, prioritizing those  
 1144 showing the strongest demographic differences. We select questions that exhibit significant variation  
 1145 across demographic groups to ensure the benchmark captures meaningful differences in responses.  
 1146 For each selected question, we create multiple test cases by pairing it with different values of the  
 1147 grouping variables (e.g., age = “18-29”, age = “30-49”). This process results in 6,343 test cases that  
 1148 specifically measure LLMs’ ability to simulate responses from narrower participant groups based on  
 1149 specified demographic characteristics. Table 4 provides a summary of the sampling process across all  
 1150 datasets, showing the minimum group size thresholds and the number of test cases in each benchmark  
 1151 split.

1152 Table 4: Dataset Sampling Summary; NaN refers to dataset that is only available in aggregated form  
 1153 and no grouping size is known.

1155 <b>Dataset</b>	1156 <b>Min. Group</b>	1157 <b>SimBench</b>	1158 <b>SimBenchPop</b>	1159 <b>SimBenchGrouped</b>
1160 WisdomOfCrowds	1161 100	1162 1,604	1163 114	1164 –
1165 Jester	1166 100	1167 136	1168 136	1169 –
1170 Choices13k	1171 NaN	1172 14,568	1173 500	1174 –
1175 OpinionQA	1176 300	1177 1,074,392	1178 500	1179 984
1180 MoralMachineClassic	1181 100	1182 3,441	1183 15	1184 –
1185 MoralMachine	1186 100	1187 20,771	1188 500	1189 –
1190 ChaosNLI	1191 100	1192 4,645	1193 500	1194 –
1195 ESS	1196 300	1197 2,783,780	1198 500	1199 1,643
1200 Afrobarometer	1201 300	1202 517,453	1203 500	1204 1,531
1205 OSPsychBig5	1206 300	1207 1,950	1208 250	1209 –
1210 OSPsychMACH	1211 300	1212 3,682,700	1213 100	1214 –
1215 OSPsychMGKT	1216 300	1217 20,610	1218 500	1219 –
1220 OSPsychRWAS	1221 300	1222 975,585	1223 22	1224 –
1225 ISSP	1226 300	1227 594,336	1228 500	1229 940
1230 LatinoBarometro	1231 300	1232 80,684	1233 500	1234 1,245
1235 GlobalOpinionQA	1236 NaN	1237 46,329	1238 500	1239 –
1240 DICES	1241 10	1242 918,064	1243 500	1244 –
1245 NumberGame	1246 10	1247 15,984	1248 500	1249 –
1250 ConspiracyCorr	1251 300	1252 968	1253 45	1254 –
1255 TISP	1256 300	1257 172,271	1258 485	1259 –
<b>1260 Total</b>		<b>1261 10,930,271</b>	<b>1262 7,167</b>	<b>1263 6,343</b>

## 1179 D IMPLEMENTATION DETAILS

1181 For base models, we use HuggingFace Transformers (Wolf et al., 2020) to run inference on a  
 1182 single NVIDIA RTX A6000 Ada GPU. We structure prompts so that the next token corresponds  
 1183 to the model’s answer choices. For models smaller than 70B parameters, we use 8-bit quantization  
 1184 implemented in bitsandbytes (Dettmers et al., 2022), while 70B models use 4-bit quantization.

1185 For instruction-tuned models, we use API calls. OpenAI models are accessed directly through their  
 1186 API, while other models are accessed via OpenRouter. We request verbalized probability outputs in  
 1187 JSON format with temperature initially set to 0. If parsing fails, we increase temperature to 1 and

1188 retry up to 5 times. All models successfully produced valid JSON under these conditions. When  
 1189 probability outputs do not sum to 1, we apply normalization.  
 1190

1191 Our evaluation includes a diverse set of models: Qwen 2.5 (Yang et al., 2024) (0.5B-72B), Gemma 3  
 1192 (Team et al., 2025) PT and IT (4B-27B), o4-mini (OpenAI, 2025b), Claude 3.7 Sonnet (Anthropic,  
 1193 2025), DeepSeek R1 (Guo et al., 2025), DeepSeek-V3-0324 DeepSeek-AI (2024), GPT-4.1 OpenAI  
 1194 (2025a), and Llama-3.1-Instruct (8B-405B) (Meta AI, 2024).

1195 To ensure the validity of our results, we perform two checks: 1) We verify that base models assign  
 1196 the vast majority of probability mass to the provided answer options. Even for small models like  
 1197 Qwen2.5-0.5B, the sum of probabilities across answer tokens is as high as 0.98, confirming that  
 1198 models rarely predict tokens outside the designated answer space. 2) We also evaluate the effect of  
 1199 quantization on model performance using a subset of SimBench. As shown in Table 5, performance  
 1200 remains consistent across quantization levels, with minimal variation in total variation scores even for  
 1201 quantization-sensitive models like Llama-3.1.

1202 We detail below the prompts used in our experimental conditions for token probability and verbalized  
 1203 distribution prediction.

1204 The following system prompt was consistent across all experimental conditions:

1205  
 1206 You are a group of individuals with these shared characteristics:  
 1207 {default system prompt} {grouping system prompt (if any)}  
 1208

1209 For token probability prediction, we adapted the prompt structure from Nori et al. (2023):  
 1210

1211 **\*\*Question\*\*:** {question}  
 1212 Do not provide any explanation, only answer with one of the following options: {answer options}.  
 1213 **\*\*Answer\*\*:** (

1214  
 1215 Prompt for eliciting verbalized probability prediction:

1216  
 1217 **\*\*Question\*\*:** {question}  
 1218 Estimate what percentage of your group would choose each option. Follow these rules:  
 1219 1. Use whole numbers from 0 to 100  
 1220 2. Ensure the percentages sum to exactly 100  
 1221 3. Only include the numbers (no % symbols)  
 1222 4. Use this exact valid JSON format: {answer options} and do NOT include anything else.  
 1223 5. Only output your final answer and nothing else. No explanations or intermediate steps are  
 1224 → needed.  
 1225 Replace X with your estimated percentages for each option.  
 1226 **\*\*Answer\*\*:**

1227  
 1228 Prompt for zero-shot CoT:

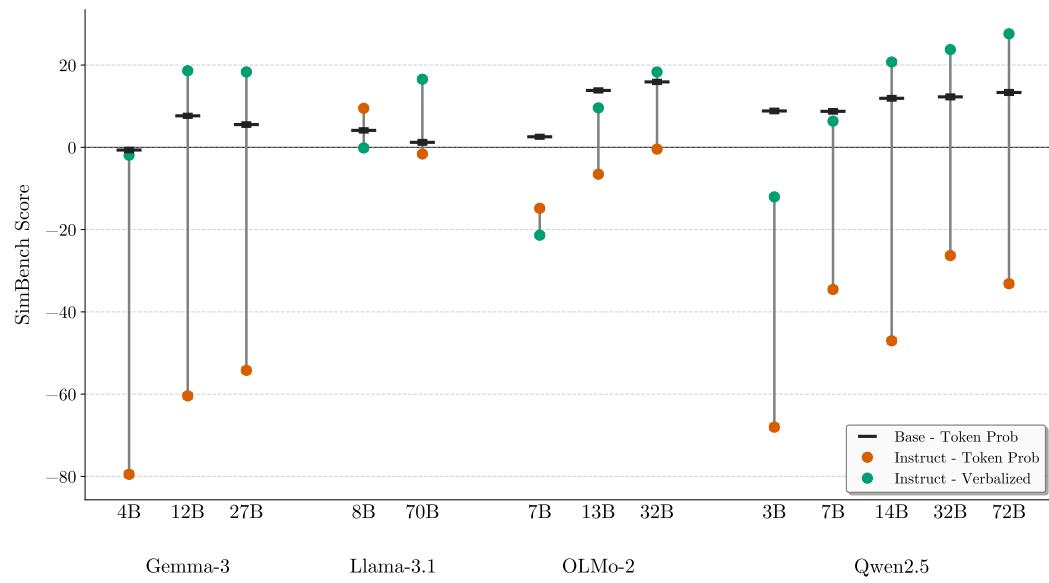
1229  
 1230 **\*\*Question\*\*:** {question}  
 1231 Estimate what percentage of your group would choose each option.  
 1232 Think step by step about how people with your shared characteristics would reason about this  
 1233 → question.  
 1234 Consider different perspectives within your group and what factors would influence their choices.  
 1235 Please provide your reasoning first, then give your final answer in JSON format.  
 1236 Follow these rules for your final answer:  
 1237 1. Use whole numbers from 0 to 100  
 1238 2. Ensure the percentages sum to exactly 100  
 1239 3. Only include the numbers (no % symbols)  
 1240 4. Use this exact valid JSON format: {json\_format\_str}  
 1241 5. Replace X with your estimated percentages for each option.  
 1242 **\*\*Answer\*\*:**

## 1242 E VALIDATION OF ELICITATION METHOD

1245 A key methodological choice in SIMBENCH is how to elicit probability distributions from LLMs. For  
 1246 base models, we use direct token probabilities from the first token of the response. For instruction-  
 1247 tuned models, however, two primary methods exist: direct token probabilities and requesting a  
 1248 “verbalized” distribution (e.g., a JSON object with percentages). To validate our choice of using  
 1249 verbalized distributions for instruction-tuned models, as recommended by recent work (Tian et al.,  
 1250 2023; Meister et al., 2025), we conducted a direct comparison.

1251 Figure 5 compares the SIMBENCH scores for several instruction-tuned models using both methods.  
 1252 The results are unequivocal: using verbalized distributions (teal dots) dramatically and consistently  
 1253 outperforms direct token probabilities (orange dots) for every instruction-tuned model tested. In  
 1254 many cases, using token probabilities results in scores far below zero, indicating that the model’s raw  
 1255 logits are poorly calibrated for this task after instruction-tuning. In contrast, base models (black bars)  
 1256 perform reasonably well with token probabilities, as they are not subject to the same post-training  
 1257 shifts.

1258 This analysis provides strong empirical support for our methodological decision to use token proba-  
 1259 bilities for base models and verbalized distributions for instruction-tuned models, ensuring that we  
 1260 are evaluating each model class using the most effective and well-calibrated elicitation technique.



1281 **Figure 5: Verbalized distributions are superior for instruction-tuned models.** This plot compares  
 1282 SIMBENCH scores for base models (using token probabilities) against their instruction-tuned counter-  
 1283 parts using either token probabilities (orange) or verbalized distributions (teal). The large vertical gap  
 1284 for all instruction-tuned models demonstrates the significant performance gain from using verbalized  
 1285 distributions, validating our choice of elicitation method.

1286  
 1287  
 1288  
 1289 **Table 5: Total Variation for different models at various quantization levels. Lower values indicate**  
 1290 **better performance.**

Model	4-bit	8-bit	16-bit	32-bit
Llama-3.1-8B-Instruct	0.272	0.266	0.262	0.262
Qwen2.5-7B	0.307	0.307	0.306	0.307

1296 **F METRIC ROBUSTNESS CHECK**  
1297

1298 TVD ranges from 0 (perfect match) to 1 (complete disagreement), with lower values indicating  
 1299 better simulation fidelity. TVD provides an interpretable measure of how closely model predictions  
 1300 align with actual human response distributions. TVD is particularly well-suited for simulation  
 1301 evaluation compared to alternatives like KL divergence or Jensen-Shannon divergence (JSD). Unlike  
 1302 KL divergence, TVD remains well-defined even when the model assigns zero probability to responses  
 1303 that humans give, avoiding the infinite penalties that KL would impose in such cases. Additionally,  
 1304 TVD is symmetric and bounded, making it more interpretable across different datasets and response  
 1305 distributions than KL divergence. While JSD offers similar advantages in terms of symmetry and  
 1306 boundedness, TVD provides a more direct and intuitive interpretation of the maximum possible error  
 1307 in probability estimates. This property is especially valuable when evaluating how accurately models  
 1308 simulate the distribution of human responses rather than just matching the most likely response. For  
 1309 further discussion on TVD as an evaluation metric, see also Meister et al. (2025). We show the results  
 1310 of Table 1 in terms of raw TVD values in Table 10.

1311 To ensure our findings are robust across different metrics, we complement TVD with two alternative  
 1312 metrics: Jensen-Shannon Divergence (JSD) and Spearman’s Rank Correlation (RC). Table 6 presents  
 1313 these metrics for a subset of evaluated models. The strong Pearson correlation between TVD and  
 1314 JSD ( $r = 0.92$ ) indicates these metrics provide consistent model rankings. The moderate negative  
 1315 correlation ( $r = -0.57$ ) between TVD and RC is expected, as lower distances correspond to higher  
 1316 correlations. This multi-metric evaluation confirms that our model comparisons remain consistent  
 1317 across different statistical measures.

1318 We chose the uniform distribution as the primary baseline  $U$  because it represents a state of maximum  
 1319 uncertainty, or the “zero-knowledge” guess. This provides the most conservative and universally  
 1320 applicable baseline across questions with varying numbers of choices. While other baselines, such as  
 1321 a majority-class baseline, could be considered, they would incorporate some knowledge about the  
 1322 distribution, making the  $S=0$  point less interpretable as a true “no-skill” score.

1323 Table 6: Comparison of models on three metrics: Total Variation Distance (TVD), Jensen-Shannon  
 1324 Divergence (JSD), and Spearman Rank Correlation (RC). Lower values are better for TVD and JSD;  
 1325 higher is better for RC.

1327 <b>Model</b>	1328 <b>Total Variation</b>	1329 <b>JS Divergence</b>	1330 <b>Rank Correlation</b>
1329 Claude-3.7-Sonnet	0.191	0.057	0.673
1330 Claude-3.7-Sonnet-4000	0.195	0.060	0.648
1331 DeepSeek-R1	0.211	0.069	0.623
1332 DeepSeek-V3-0324	0.216	0.069	0.620
1333 GPT-4.1	0.209	0.070	0.646
1334 Llama-3.1-405B-Instruct	0.231	0.085	0.593
1335 o4-mini-high	0.225	0.079	0.621
1336 o4-mini-low	0.230	0.082	0.609

1339 **G FULL SIMBENCH RESULTS (RQ1)**  
1340

1341 We show the SIMBENCH scores for all the 45 models we evaluate in Table 7. We show the scaling  
 1342 law plots for all models in Figure 6.

1345 **H STATISTICAL ANALYSIS OF MODEL PERFORMANCE**  
1346

1348 To ensure the robustness of our findings, we conducted a comprehensive statistical analysis of  
 1349 model performance on SIMBENCH. This includes confidence intervals for overall scores, pairwise  
 significance tests between models, and within-family analyses to validate our scaling law observations.

1350  
 1351  
 1352  
 1353 **Table 7: Overall simulation ability** as measured by SIMBENCH score  $S$  averaged across the two  
 1354 main splits of SIMBENCH. Reasoning models are highlighted in *italics*. Models are sorted by score.  
 1355 Models below the dotted line perform worse than a uniform baseline.

Model	Type	Release	$S (\uparrow)$
Claude-3.7-Sonnet	Instr.	Closed	40.80
<i>Claude-3.7-Sonnet-4000</i>	Instr.	Closed	39.46
GPT-4.1	Instr.	Closed	34.55
<i>DeepSeek-R1</i>	Instr.	Open	34.52
<i>DeepSeek-V3-0324</i>	Instr.	Open	32.89
<i>o4-mini-high</i>	Instr.	Closed	28.99
Llama-3.1-405B-Instruct	Instr.	Open	28.40
<i>o4-mini-low</i>	Instr.	Closed	27.77
Qwen2.5-72B-Instruct	Instr.	Open	27.61
Qwen2.5-32B-Instruct	Instr.	Open	23.76
Qwen2.5-14B-Instruct	Instr.	Open	20.75
OLMo-2-0325-32B-DPO	Instr.	Open	19.80
Gemma-3-12B-IT	Instr.	Open	18.62
Gemma-3-27B-IT	Instr.	Open	18.33
OLMo-2-0325-32B-Instruct	Instr.	Open	18.32
Llama-3.1-70B-Instruct	Instr.	Open	16.56
OLMo-2-0325-32B	Base	Open	15.90
Llama-3.1-Minitaur-8B	Base	Open	14.50
OLMo-2-1124-13B	Base	Open	13.83
Qwen2.5-72B	Base	Open	13.34
Qwen2.5-32B	Base	Open	12.27
Qwen2.5-14B	Base	Open	11.92
OLMo-2-0325-32B-SFT	Instr.	Open	11.28
OLMo-2-1124-13B-Instruct	Instr.	Open	9.59
OLMo-2-1124-13B-DPO	Instr.	Open	9.42
Qwen2.5-3B	Base	Open	8.84
Qwen2.5-7B	Base	Open	8.75
Llama-3.1-Centaur-70B	Base	Open	8.54
Gemma-3-12B-PT	Base	Open	7.66
Qwen2.5-7B-Instruct	Instr.	Open	6.36
Gemma-3-27B-PT	Base	Open	5.53
Qwen2.5-1.5B	Base	Open	5.34
Llama-3.1-8B	Base	Open	4.12
OLMo-2-1124-7B	Base	Open	2.56
Llama-3.1-70B	Base	Open	1.21
Llama-3.1-8B-Instruct	Instr.	Open	-0.15
Gemma-3-4B-PT	Base	Open	-0.65
Gemma-3-4B-IT	Instr.	Open	-1.91
Qwen2.5-0.5B	Base	Open	-3.00
OLMo-2-1124-7B-SFT	Instr.	Open	-11.36
Qwen2.5-3B-Instruct	Instr.	Open	-12.04
Gemma-3-1B-PT	Base	Open	-16.17
OLMo-2-1124-7B-DPO	Instr.	Open	-19.62
OLMo-2-1124-13B-SFT	Instr.	Open	-20.54
OLMo-2-1124-7B-Instruct	Instr.	Open	-21.36

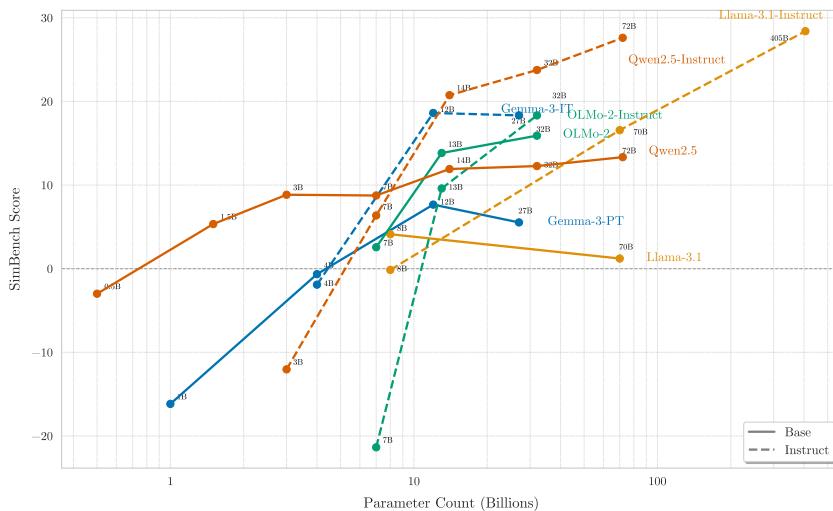


Figure 6: **Model parameter count vs. simulation ability.** We measure model size by parameter count and simulation ability by SIMBENCH score  $S$  averaged across the two main splits of SIMBENCH.

## H.1 OVERALL MODEL RANKING AND SIGNIFICANCE

Table 8 expands on the results from Table 1 in the main paper. We report the 95% confidence interval (CI) for each model's mean SIMBENCH score. We also perform independent two-sample t-tests to determine if the performance difference between each model and the one ranked immediately below it is statistically significant. The results confirm a clear and statistically robust performance hierarchy among the evaluated models.

Table 8: Detailed statistical analysis of SIMBENCH scores for the top 8 models. Pairwise significance tests compare each model to the one ranked immediately below it. Symbols indicate significance: \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ .

Rank	Model	Score (S) $\pm$ SE	95% CI
1	Claude-3.7-Sonnet	40.80(43)*	[39.96, 41.64]
2	Claude-3.7-Sonnet-4000	39.46(42)***	[38.63, 40.29]
3	GPT-4.1	34.56(48)	[33.62, 35.50]
4	DeepSeek-R1	34.52(43)**	[33.68, 35.37]
5	DeepSeek-V3-0324	32.90(42)***	[32.07, 33.72]
6	o4-mini-high	28.99(52)	[27.97, 30.02]
7	Llama-3.1-405B-Instruct	28.41(49)	[27.46, 29.36]
8	o4-mini-low	27.77(51)***	[26.77, 28.78]

## H.2 WITHIN-FAMILY PERFORMANCE ANALYSIS

To investigate the impact of model size (RQ2), we conducted one-way ANOVA followed by post-hoc Tukey HSD tests for pairwise comparisons within model families. The results, presented in Table 9, show that for most families, increases in model size lead to statistically significant improvements in simulation ability, supporting the [scaling trends](#) observed in Figure 2.

1458 Table 9: Within-family pairwise performance comparisons. All differences are statistically significant  
 1459 at  $p < 0.001$  unless otherwise noted (ns).

Model Family	Comparison	Result
<b>Llama-3.1-Instruct</b> (ANOVA: $F=518.8$ , $p < 0.001$ )	405B vs 70B	$\Delta S = 11.84$ ( $p < 0.001$ )
	405B vs 8B	$\Delta S = 28.55$ ( $p < 0.001$ )
	70B vs 8B	$\Delta S = 16.71$ ( $p < 0.001$ )
<b>Gemma-3-IT</b> (ANOVA: $F=710.1$ , $p < 0.001$ )	27B vs 12B	$\Delta S = 0.29$ (ns, $p = 0.673$ )
	12B vs 4B	$\Delta S = 20.54$ ( $p < 0.001$ )
	27B vs 4B	$\Delta S = 20.24$ ( $p < 0.001$ )
<b>Qwen2.5-Instruct</b> (ANOVA: $F=518.8$ , $p < 0.001$ )	72B vs 32B	$\Delta S = 3.85$ ( $p < 0.001$ )
	72B vs 14B	$\Delta S = 6.86$ ( $p < 0.001$ )
	72B vs 7B	$\Delta S = 21.24$ ( $p < 0.001$ )
	32B vs 14B	$\Delta S = 3.01$ ( $p < 0.001$ )
	32B vs 7B	$\Delta S = 17.39$ ( $p < 0.001$ )
	14B vs 7B	$\Delta S = 14.38$ ( $p < 0.001$ )

## I ANALYSIS OF THE ALIGNMENT-SIMULATION TRADEOFF

### I.1 OBSERVATIONAL ANALYSIS

As a complement to the analysis in §4.4, this section provides a direct observational comparison of how base and instruction-tuned models perform as a function of human response plurality. We operationalize response plurality as the normalized entropy of the human response distribution, where high entropy indicates widespread disagreement while a low entropy indicates strong consensus. Simulation fidelity is measured by Total Variation Distance (TVD), where lower values indicate better performance. Figure 7 visualizes this relationship for all question-model pairs in SimBenchPop, separated by model type. The plots reveal a clear and divergent behavioral pattern:

- **Base models** exhibit a negative correlation between entropy and TVD. This demonstrates an observable property of this model class: they are generally more accurate (lower TVD) on high-entropy questions where human opinions are diverse.
- **Instruction-tuned models** exhibit a positive correlation. This demonstrates the opposite property: they are generally more accurate on low-entropy questions where humans have reached a consensus.

This direct visualization establishes a core empirical finding: the two model classes have fundamentally different strengths when simulating human responses across the spectrum of opinion plurality. The analysis in the main paper (§4.4) builds upon this observation to more formally quantify the effect of post-training that drives this divergence.

### I.2 REGRESSION ANALYSIS

To formally test the relationship between human response entropy and simulation performance across different model types, we fit an Ordinary Least Squares (OLS) regression model predicting Total Variation (TV) distance at the individual question-model level. The model specification was as follows:

$$\text{Total\_Variation} \sim C(\text{dataset\_name}) + C(\text{model}) + C(\text{instruct\_flag}) : \text{Human\_Normalized\_Entropy} \quad (3)$$

Here, *Total\_Variation* is the dependent variable.  $C(\text{dataset\_name})$  and  $C(\text{model})$  represent fixed effects for each dataset and model, respectively, controlling for baseline differences in difficulty and capability. The crucial term is the interaction  $C(\text{instruct\_flag}) : \text{Human\_Normalized\_Entropy}$ , where *instruct\_flag* is a binary indicator for instruction-tuned models (0 for base, 1 for instruction-tuned).

The key results from Table 11 are the coefficients for the interaction terms:

1512 Table 10: TVD for each model in SimBenchPop and SimBenchGrouped. Lower values indicate  
 1513 better performance. PT and IT refer to pretrained and instruction-tuned versions, respectively.

Model	SimBenchPop	SimBenchGrouped	Average
<i>Baselines</i>			
Uniform baseline	0.335	0.362	0.348
<i>Models</i>			
Claude-3.7-Sonnet	0.197	0.183	0.191
Claude-3.7-Sonnet-4000	0.201	0.189	0.195
GPT-4.1	0.212	0.206	0.209
DeepSeek-R1	0.211	0.212	0.211
DeepSeek-V3-0324	0.215	0.218	0.216
o4-mini-high	0.235	0.214	0.225
o4-mini-low	0.234	0.226	0.230
Llama-3.1-405B-Instruct	0.237	0.225	0.231
Qwen2.5-72B-Instruct	0.229	0.246	0.237
Qwen2.5-32B-Instruct	0.242	0.258	0.250
OLMo-2-0325-32B-DPO	0.258	0.258	0.258
Qwen2.5-14B-Instruct	0.247	0.270	0.258
OLMo-2-0325-32B-Instruct	0.263	0.260	0.261
Llama-3.1-70B-Instruct	0.277	0.247	0.263
Gemma-3-12B-IT	0.262	0.274	0.267
Gemma-3-27B-IT	0.270	0.273	0.272
OLMo-2-0325-32B	0.271	0.297	0.283
OLMo-2-0325-32B-SFT	0.298	0.265	0.283
Qwen2.5-72B	0.268	0.300	0.283
Qwen2.5-32B	0.273	0.308	0.290
Llama-3.1-Minitaur-8B	0.288	0.296	0.292
OLMo-2-1124-13B	0.284	0.302	0.293
Qwen2.5-14B	0.285	0.314	0.298
OLMo-2-1124-13B-DPO	0.293	0.306	0.299
OLMo-2-1124-13B-Instruct	0.295	0.304	0.299
Qwen2.5-7B	0.290	0.326	0.307
Llama-3.1-Centaur-70B	0.309	0.313	0.311
Qwen2.5-7B-Instruct	0.292	0.332	0.311
Qwen2.5-3B	0.300	0.327	0.313
Gemma-3-12B-PT	0.310	0.317	0.314
Gemma-3-27B-PT	0.309	0.325	0.317
Llama-3.1-8B-Instruct	0.321	0.318	0.320
Qwen2.5-1.5B	0.321	0.324	0.322
Llama-3.1-8B	0.326	0.323	0.324
Llama-3.1-70B	0.331	0.324	0.328
OLMo-2-1124-7B	0.324	0.349	0.336
Gemma-3-4B-PT	0.334	0.341	0.337
Gemma-3-4B-IT	0.337	0.341	0.339
Qwen2.5-0.5B	0.337	0.364	0.349
OLMo-2-1124-7B-SFT	0.393	0.355	0.375
Qwen2.5-3B-Instruct	0.397	0.363	0.381
Gemma-3-1B-PT	0.382	0.414	0.397
OLMo-2-1124-7B-DPO	0.413	0.382	0.399
OLMo-2-1124-7B-Instruct	0.420	0.386	0.404
OLMo-2-1124-13B-SFT	0.416	0.414	0.415

1564  
1565

1566

1567

1568

1569

1570

1571

1572

1573

1574

1575

Table 11: Results: Ordinary least squares

1576

Model:	OLS	Adj. R-squared:	0.168
Dependent Variable:	Total_Variation	AIC:	-163587.0704
Date:	2025-09-23 12:11	BIC:	-163085.0895
No. Observations:	207837	Log-Likelihood:	81843.
Df Model:	48	F-statistic:	875.3
Df Residuals:	207788	Prob (F-statistic):	0.00
R-squared:	0.168	Scale:	0.026644

1580

	Coeff.	Std.Err.	t	P> t	[0.025	0.975]
Intercept	0.1811	0.0024	75.6586	0.0000	0.1764	0.1858
C(dataset_name)[T.ChaosNLI]	-0.0372	0.0019	-19.2201	0.0000	-0.0409	-0.0334
C(dataset_name)[T.Choices13k]	-0.0972	0.0019	-49.9249	0.0000	-0.1010	-0.0934
C(dataset_name)[T.ConspiracyCorr]	-0.0240	0.0047	-5.0722	0.0000	-0.0333	-0.0147
C(dataset_name)[T.DICES]	-0.0331	0.0021	-15.8245	0.0000	-0.0372	-0.0290
C(dataset_name)[T.ESSE]	-0.0261	0.0019	-13.5145	0.0000	-0.0299	-0.0223
C(dataset_name)[T.GlobalOpinionQA]	-0.0406	0.0019	-21.1621	0.0000	-0.0444	-0.0369
C(dataset_name)[T.ISSP]	-0.0292	0.0019	-15.1813	0.0000	-0.0330	-0.0255
C(dataset_name)[T.Jester]	0.1104	0.0029	37.4254	0.0000	0.1046	0.1162
C(dataset_name)[T.LatinoBarometro]	-0.0368	0.0019	-18.9530	0.0000	-0.0406	-0.0330
C(dataset_name)[T.MoralMachine]	-0.0438	0.0019	-22.6987	0.0000	-0.0476	-0.0401
C(dataset_name)[T.MoralMachineClassic]	-0.1676	0.0079	-21.0886	0.0000	-0.1831	-0.1520
C(dataset_name)[T.NumberGame]	-0.0852	0.0019	-44.4242	0.0000	-0.0889	-0.0814
C(dataset_name)[T.OSPsychBig5]	-0.1218	0.0024	-51.0313	0.0000	-0.1265	-0.1171
C(dataset_name)[T.OSPsychMACH]	-0.0200	0.0033	-5.9606	0.0000	-0.0265	-0.0134
C(dataset_name)[T.OSPsychMGKT]	-0.1121	0.0019	-57.9995	0.0000	-0.1159	-0.1083
C(dataset_name)[T.OSPsychRWAS]	0.0105	0.0066	1.5909	0.1116	-0.0024	0.0235
C(dataset_name)[T.OpinionQA]	-0.1080	0.0019	-56.3479	0.0000	-0.1118	-0.1043
C(dataset_name)[T.TISP]	-0.0429	0.0020	-21.9962	0.0000	-0.0467	-0.0391
C(dataset_name)[T.WisdomOfCrowds]	-0.0224	0.0032	-7.0814	0.0000	-0.0286	-0.0162
C(Model)[T.DeepSeek-R1]	0.0114	0.0024	4.8093	0.0000	0.0067	0.0160
C(Model)[T.DeepSeek-V3-0324]	0.0158	0.0024	6.6887	0.0000	0.0112	0.0204
C(Model)[T.GPT-4.1]	0.0122	0.0024	5.1618	0.0000	0.0076	0.0168
C(Model)[T.Gemma-3-12B-IT]	0.0622	0.0024	26.3302	0.0000	0.0575	0.0668
C(Model)[T.Gemma-3-12B-PT]	0.3521	0.0031	115.0964	0.0000	0.3461	0.3581
C(Model)[T.Gemma-3-27B-IT]	0.0711	0.0024	30.1015	0.0000	0.0665	0.0757
C(Model)[T.Gemma-3-27B-PT]	0.3509	0.0031	114.6963	0.0000	0.3449	0.3569
C(Model)[T.Llama-3.1-405B-Instruct]	0.0373	0.0024	15.7766	0.0000	0.0326	0.0419
C(Model)[T.Llama-3.1-70B]	0.3730	0.0031	121.9329	0.0000	0.3670	0.3790
C(Model)[T.Llama-3.1-70B-Instruct]	0.0772	0.0024	32.7116	0.0000	0.0726	0.0819
C(Model)[T.Llama-3.1-8B]	0.3676	0.0031	120.1756	0.0000	0.3616	0.3736
C(Model)[T.Llama-3.1-8B-Instruct]	0.1212	0.0024	51.3289	0.0000	0.1166	0.1258
C(Model)[T.OLMo-2-0325-32B]	0.3125	0.0031	102.1507	0.0000	0.3065	0.3185
C(Model)[T.OLMo-2-0325-32B-Instruct]	0.0636	0.0024	26.9284	0.0000	0.0590	0.0682
C(Model)[T.OLMo-2-1124-13B]	0.3261	0.0031	106.5964	0.0000	0.3201	0.3321
C(Model)[T.OLMo-2-1124-13B-Instruct]	0.0953	0.0024	40.3576	0.0000	0.0907	0.0999
C(Model)[T.Qwen2.5-1.5B]	0.3624	0.0031	118.4639	0.0000	0.3564	0.3684
C(Model)[T.Qwen2.5-14B]	0.3264	0.0031	106.6913	0.0000	0.3204	0.3324
C(Model)[T.Qwen2.5-14B-Instruct]	0.0481	0.0024	20.3827	0.0000	0.0435	0.0528
C(Model)[T.Qwen2.5-32B]	0.3152	0.0031	103.0509	0.0000	0.3092	0.3212
C(Model)[T.Qwen2.5-32B-Instruct]	0.0426	0.0024	18.0362	0.0000	0.0380	0.0472
C(Model)[T.Qwen2.5-3B]	0.3422	0.0031	111.8592	0.0000	0.3362	0.3482
C(Model)[T.Qwen2.5-72B]	0.3103	0.0031	101.4261	0.0000	0.3043	0.3163
C(Model)[T.Qwen2.5-72B-Instruct]	0.0292	0.0024	12.3690	0.0000	0.0246	0.0338
C(Model)[T.Qwen2.5-7B]	0.3314	0.0031	108.3263	0.0000	0.3254	0.3374
C(Model)[T.o4-mini-high]	0.0354	0.0024	15.0086	0.0000	0.0308	0.0401
C(Model)[T.o4-mini-low]	0.0344	0.0024	14.5682	0.0000	0.0298	0.0390
C(instruct_flag)[base]Human_Normalized_Entropy	-0.2451	0.0024	-100.0564	0.0000	-0.2499	-0.2403
C(instruct_flag)[instruct]Human_Normalized_Entropy	0.0997	0.0022	46.1412	0.0000	0.0955	0.1039

1610

Omnibus:	32553.497	Durbin-Watson:	1.732
Prob(Omnibus):	0.000	Jarque-Bera (JB):	60320.484
Skew:	0.995	Prob(JB):	0.000
Kurtosis:	4.733	Condition No.:	30

1611

1612

1613

1614

1615

1616

1617

1618

1619

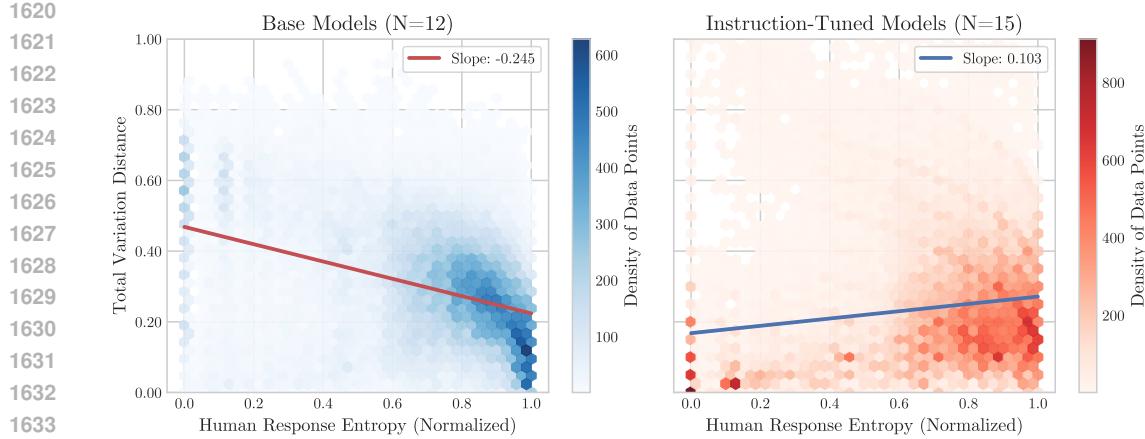


Figure 7: **Response plurality vs. simulation fidelity** for base and instruction-tuned models on all questions in SimBenchPop. We measure response plurality by normalized entropy of the human response distribution and simulation fidelity by total variation distance at the question level.

- For base models: The coefficient on the interaction between base models and Human Normalized Entropy is  $-0.2451$  ( $p < 0.001$ ), indicating that for every one-unit increase in normalized entropy, the TVD decreases by approximately 0.25 units. This means that base models perform *better* (lower TVD) when simulating human populations with more diverse opinions.
- For instruction-tuned models: The coefficient on the interaction between instruction-tuned models and Human Normalized Entropy is  $+0.0997$  ( $p < 0.001$ ), indicating that for every one-unit increase in normalized entropy, the TVD increases by approximately 0.11 units. This means that instruction-tuned models perform *worse* (higher TVD) when simulating human populations with more diverse opinions.

These coefficients are both highly statistically significant ( $p < 0.001$ ) and represent substantial effect sizes given that TVD ranges from 0 to 1. The model as a whole explains approximately 17% of the variance in TVD ( $R^2 = 0.168$ ), which is substantial for a dataset of this size and complexity. The opposite signs of these coefficients provide strong evidence for our hypothesis that base models and instruction-tuned models respond differently to the challenge of simulating populations with diverse opinions. This pattern holds even after controlling for the specific datasets and models involved, suggesting it represents a general property of the two model classes rather than an artifact of particular model or evaluation datasets.

### I.3 CAUSAL MEDIATION ANALYSIS: DECOMPOSING THE DUAL EFFECTS OF INSTRUCTION TUNING

To formally test the causal mechanisms behind the alignment-simulation tradeoff (§4.4), we conducted a causal mediation analysis. This analysis aims to decompose the effect of instruction tuning, separating its impact on the model’s output entropy from other effects like improved instruction following.

**Causal Model** Our analysis is based on a linear mediation model designed to decompose the total effect of instruction tuning into direct and indirect pathways. The hypothesized causal graph is as follows:

Instruction Tuning (Treatment,  $X$ )  $\rightarrow$  Model Prediction Entropy (Mediator,  $M$ )  $\rightarrow$  SimBench Score (Outcome,  $Y$ )

The central hypothesis is that instruction tuning ( $X$ ) affects simulation performance ( $Y$ ) at least partially *through* its systematic effect on the entropy of the model’s output distribution ( $M$ ).

1674  
 1675     **Methodology and Model Specification** The analysis was implemented in Python using the  
 1676     statsmodels library (Seabold & Perktold, 2010). We fit two Ordinary Least Squares (OLS)  
 1677     regression models to estimate the relevant causal paths, following the standard mediation framework:

1678     1. **Mediator Model (Path  $a$ ):** We first model the effect of instruction tuning on the mediator (model  
 1679     prediction entropy). The model is specified as:

$$M_i = \alpha_M + aX_i + \mathbf{\Gamma}_M^T \mathbf{Z}_i + \epsilon_{M,i} \quad (4)$$

1680  
 1681     1682 where  $M_i$  is the normalized entropy of the model’s prediction for a given question,  $X_i$  is a binary  
 1683     1684 indicator for whether the model is instruction-tuned (1 if true, 0 if base model), and  $\mathbf{Z}_i$  is a vector  
 1685     1686 of control variables. The coefficient  $a$  captures the average effect of instruction tuning on model  
 1687     1688 prediction entropy.

1689     2. **Outcome Model (Paths  $b$  and  $c'$ ):** We then model the outcome (SimBench Score) as a function  
 1690     1691 of both the treatment and the mediator:

$$Y_i = \alpha_Y + c'X_i + bM_i + \mathbf{\Gamma}_Y^T \mathbf{Z}_i + \epsilon_{Y,i} \quad (5)$$

1692  
 1693     1694 where  $Y_i$  is the SIMBENCH Score. The coefficient  $c'$  represents the *direct effect* of instruction tuning  
 1695     1696 on performance, holding model entropy constant. The coefficient  $b$  represents the effect of model  
 1697     1698 entropy on performance, holding instruction tuning constant.

1699     **Control Variables** In both models, the vector  $\mathbf{Z}_i$  includes a set of control variables to account for  
 1700     1701 potential confounders:

- **Human Response Entropy:** We control for the normalized entropy of the ground-truth human answer distribution to isolate the model’s behavior from the inherent plurality of the question.
- **Fixed Effects:** We include fixed effects for both the model family (e.g., Llama-3.1, Qwen2.5) and the dataset to absorb any baseline differences in performance or entropy across these groups.

1702     **Effect Calculation** The key effects are calculated from the estimated coefficients:

- **Direct Effect ( $c'$ ):** Directly estimated from the outcome model.
- **Indirect Effect ( $a \times b$ ):** The effect of instruction tuning that is mediated through model entropy, calculated as the product of the coefficients from the two models. The statistical significance of this indirect effect was assessed using the Sobel test approximation for the standard error.
- **Total Effect:** The sum of the direct and indirect effects ( $c' + a \times b$ ).

1703  
 1704     1705 This decomposition allows us to quantify how much of instruction tuning’s overall impact on  
 1706     1707 simulation fidelity is attributable to its entropy-suppressing nature versus other factors like improved  
 1708     1709 instruction following.

1710     **Results and Interpretation** Our analysis reveals that instruction tuning has two distinct and  
 1711     1712 opposing effects on simulation ability. The overall total effect is a modest but significant improvement  
 1713     1714 of **+4.72** points on the SIMBENCH score ( $p < .001$ ). However, this net effect masks two powerful  
 1715     1716 underlying mechanisms:

1. **A Harmful Indirect Effect (-1.74 points):** Instruction tuning significantly reduces model prediction entropy (Path A:  $\beta = -0.11, p < .001$ ). In our models, higher entropy is generally associated with better performance (Path B:  $\beta = 15.60, p < .001$ ). The indirect effect ( $A \times B$ ) is therefore negative (-1.74), quantifying the performance penalty that instruction tuning imposes by forcing the model into a low-entropy, mode-seeking behavior.
2. **A Strong, Helpful Direct Effect (+6.46 points):** After accounting for the change in entropy, a large positive **direct effect** remains ( $\beta = +6.46, p < .001$ ). This reflects the benefits of instruction tuning that are independent of its impact on output diversity, such as improved instruction following and a better ability to reason about the specified persona.

1724  
 1725     **Conclusion** These results provide strong evidence for *inconsistent mediation* and resolve a key  
 1726     1727 paradox in our findings. While our analysis in §4.4 shows that instruction-tuning harms simulation  
 1728     1729 fidelity on high-entropy questions, our main leaderboard (Table 4.1) shows that the best overall  
 1730     1731 simulators are instruction-tuned. This mediation analysis explains why: the total effect of instruction

tuning is the net outcome of two larger, opposing forces. First, a **direct positive effect (+6.46 points)** on capability, likely from improved instruction- and persona-following. Second, a smaller but significant **indirect negative effect (-1.74 points)** caused by entropy suppression. The net positive effect (+4.72 points) demonstrates that, on average, the direct benefits of alignment currently outweigh the harm from reduced distributional diversity. Future work on creating SOTA simulators should therefore focus on developing hybrid or “distribution-preserving” alignment methods that retain the direct benefits of instruction-tuning while mitigating its harmful, entropy-reducing side effects.

## J DETAILED CORRELATION ANALYSIS (RQ6)

To support our analysis in Section 4.6, this appendix provides the detailed data sources and scatter plots illustrating the correlation between SIMBENCH scores and five external capability benchmarks. This analysis includes the subset of our evaluated models for which performance data on these external benchmarks could be reliably sourced. The benchmark performance data was collected from model developers’ technical reports, the Open LLM Leaderboard Fourrier et al. (2024), and Vals AI, Inc. (2025). Table 12 summarizes the Pearson correlation coefficients, and Figure 8 presents the individual scatter plots for each benchmark.

Table 12: Summary of Pearson Correlation Coefficients ( $r$ ) between SIMBENCH scores and external capability benchmarks for the models evaluated in our study.

Capability Benchmark	Pearson’s $r$
MMLU-Pro	0.939
GPQA Diamond	0.862
IF-Eval	0.786
Chatbot Arena ELO	0.708
OTIS AIME	0.479

## K CASE STUDY: DETAILED ANALYSIS OF CENTAUR

We present a detailed visualization of model performance across the spectrum of human response entropy. Figure 9 breaks down the SIMBENCH Score for the Llama-3.1 8B and 70B models: base, instruction-tuned, and specialist cognitive-tuned (Minitaur/Centaur), binned by the normalized entropy of the human ground truth. The figure reveals that these two fine-tuning paradigms improve simulation ability in fundamentally different and complementary ways. **General-purpose instruction tuning** excels in low-entropy regimes where there is a clear human consensus. The orange, dashed lines for both 8B and 70B Instruct models show the highest performance (SIMBENCH Score) when entropy is low, but this advantage systematically decays as human opinions become more diverse. This aligns with its mode-seeking objective: it trains the model to identify and produce a single “correct” or preferred response. **Specialist cognitive tuning**, in contrast, mirrors the behavior of base models. The green and blue dash-dotted lines for Minitaur and Centaur show a distinct pattern: performance is weaker on low-entropy tasks but progressively improves as human response entropy increases. This suggests that fine-tuning on behavioral data preserves or even enhances the base model’s mass-covering ability to represent a diverse distribution of outcomes, rather than forcing it into a single mode.

This qualitative divergence is key. The two methods are not just different in degree, but in kind. Instruction-tuning boosts performance by sharpening a model’s ability to follow prompt instructions and converge on a consensus answer. Specialist tuning boosts performance by aligning the model’s internal representations more closely with the patterns of human choice. Because they target different mechanisms, their benefits are not mutually exclusive. This suggests that perhaps future gains in LLM simulation will come from hybrid approaches that synthesize both paradigms, creating models that are both generally capable and foundationally aligned with the nuances of human behavior.

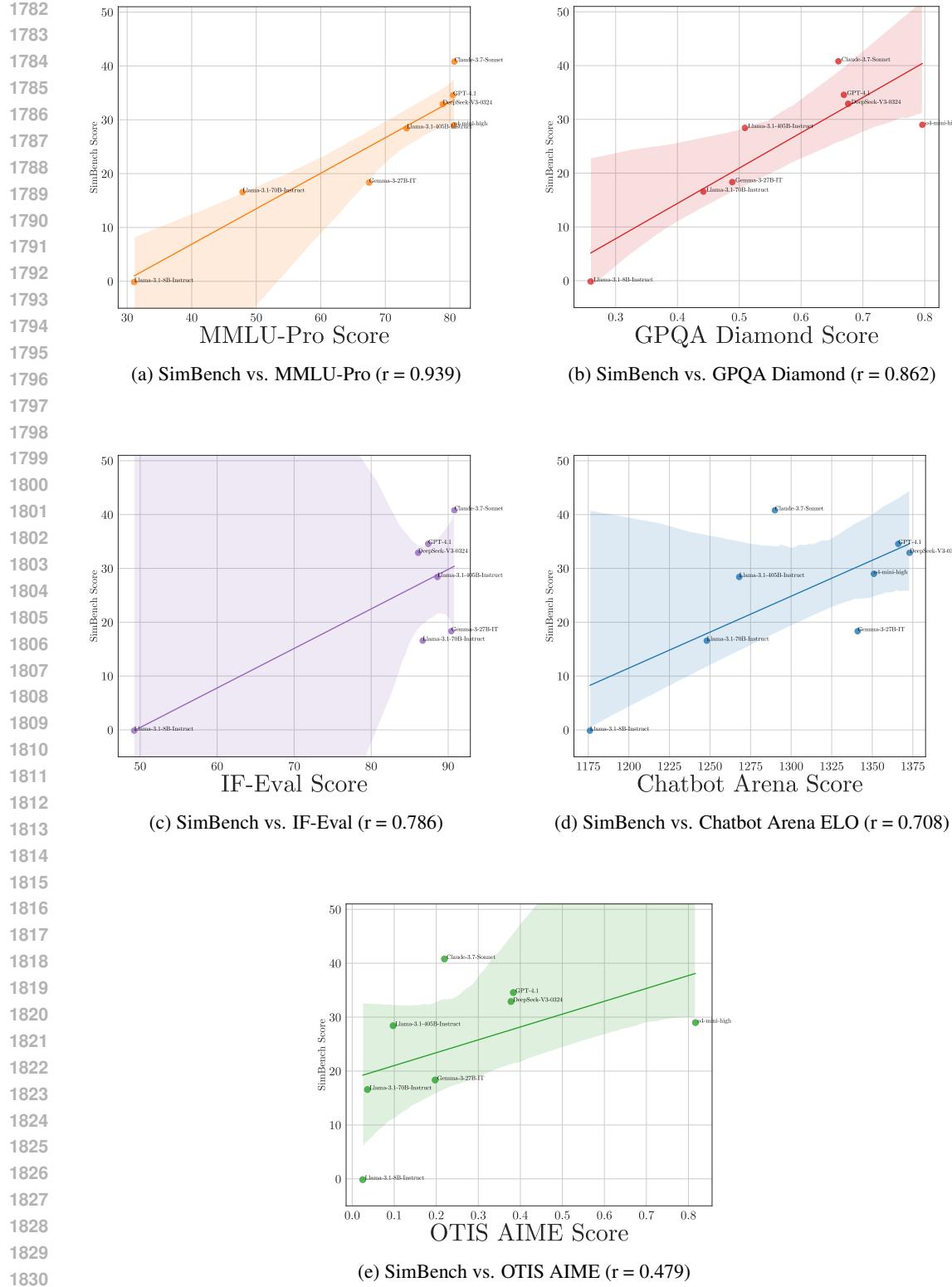


Figure 8: Scatter plots showing the correlation between average SIMBENCH scores and performance on five external benchmarks. Each point represents an LLM. The strong positive correlation is most pronounced for knowledge-intensive reasoning tasks (a, b).

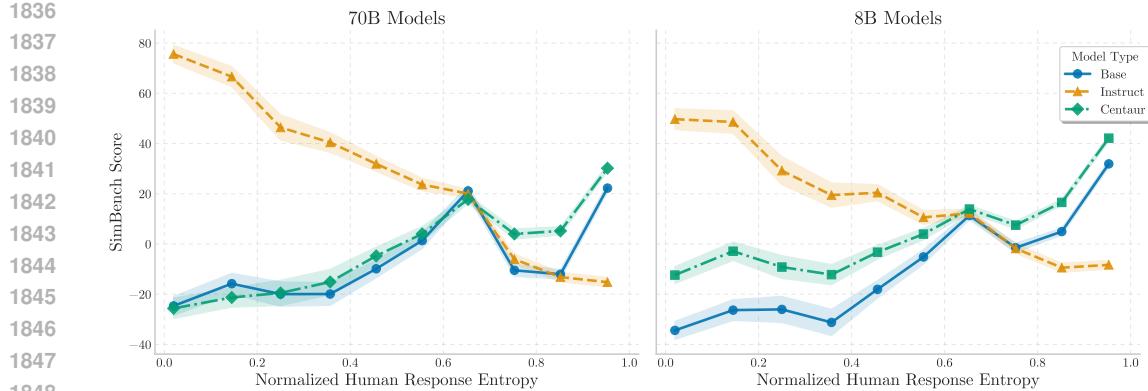


Figure 9: **Effect of Centaur Fine-Tuning.** The plots show the binned SIMBENCH Score against normalized human response entropy for Llama-3.1 models at the 70B (left) and 8B (right) scales. Shaded areas represent the 95% confidence interval of the mean score in each bin.

## L DATASET DETAILS

We provide details on each of the 20 datasets in SIMBENCH. Note that for many datasets we use only a subset of questions and participants for SIMBENCH, as a result of our preprocessing steps (§2.3).

### L.1 WISDOMOFCROWDS

**Description:** This dataset contains **factual questions** that were administered to a large number of US-based Amazon Mechanical Turk workers. The data was originally collected to study wisdom of the crowd effects.

**Questions:** 113, with an average of 518 responses per question.

**Example question:**

An analogy compares the relationship between two things or ideas to highlight some point of similarity. You will be given pairs of words bearing a relationship, and asked to select another pair of words that illustrate a similar relationship.

Which pair of words has the same relationship as 'Letter : Word'?

- (A): Page : Book
- (B): Product : Factory
- (C): Club : People
- (D): Home work : School

**Participants:** 722 US-based Amazon Mechanical Turk workers.

**Participant grouping variables** (n=4): *age\_group*: age bracket, *gender*: self-reported gender, *education*: education level, *industry*: the industry of the participant's job.

**Default System Prompt:**

You are an Amazon Mechanical Turk worker from the United States.

**License:** MIT

**Publication:** Simoiu et al. (2019)

1890 L.2 JESTER

1891

1892 **Description:** This dataset contains **jokes** for which participants provided **subjective judgments** of  
1893 how funny they found them. The data was originally collected to enable recommender systems and  
1894 collaborative filtering research.

1895

1896 **Questions:** 136, with an average of 779 responses per question.1897 **Example question:**

1898

1899 How funny is the following joke, on a scale of -10 to 10? (-10: not funny, 10: very funny)

1900

1901 How many feminists does it take to screw in a light bulb? That's not funny.

1902

1903 Options:

1904 (A): 7 to 10  
1905 (B): 3 to 6  
1906 (C): -2 to 2  
1907 (D): -5 to -3  
1908 (E): -10 to -61909 **Participants:** 7,669 volunteer participants (sociodemographics unknown) who chose to use the Jester  
1910 joke recommender website.

1911

1912 **Participant grouping variables:** None. **Default System Prompt:**

1913

1914 Jester is a joke recommender system developed at UC Berkeley to study social information filtering.  
1915 You are a user of Jester.

1916

1917 **License:** "Freely available for research use when cited appropriately."

1918

1919 **Publication:** Goldberg et al. (2001)

1920

1921

1922 L.3 CHOICES13K

1923

1924 **Description:** This dataset contains a large number of automatically generated **decision-making**  
1925 **scenarios** that present participants with two lotteries to choose from. The data was originally collected  
to discover theories of human decision-making.

1926

1927 **Questions:** 14,568, with an average of 17 responses per question.

1928

1929 **Example question:**

1930

1931

1932

1933

1934

1935

1936

1937

1938 There are two gambling machines, A and B. You need to make a choice between the machines with  
1939 the goal of maximizing the amount of dollars received. You will get one reward from the machine  
1940 that you choose. A fixed proportion of 10% of this value will be paid to you as a performance  
1941 bonus. If the reward is negative, your bonus is set to \$0.

1942

1943 Machine A: \$-1.0 with 5.0% chance, \$26.0 with 95.0% chance.

Machine B: \$21.0 with 95.0% chance, \$23.0 with 5.0% chance.

Which machine do you choose?

1944

1945

1946

1947

1948 **Participants:** 14,711 US-based Amazon Mechanical Turk workers.

1949

1950

1951 **Participant grouping variables:** None.

1952

1953 **Default System Prompt:**

1954

1955 You are an Amazon Mechanical Turk worker based in the United States.

1944 **License:** “All data are available to the public without registration at  
 1945 [github.com/jcpeterson/choices13k](https://github.com/jcpeterson/choices13k)”.

1946  
 1947 **Publication:** Peterson et al. (2021)

1948

1949 L.4 OPINIONQA

1950

1951 **Description:**

1952 This dataset contains **survey questions** that ask participants to provide **self-assessments** and **subjective judgments**. The data was sourced from the Pew Research American Trends Panel, and then  
 1953 repurposed to evaluate LLM alignment with the opinions of different sociodemographic groups.

1954  
 1955 **Questions:** 736, with an average of 5,339 responses per question.

1956  
 1957 **Example question:**

1958

1959 How would you describe your household’s financial situation?

1960

1961 (A): Live comfortably  
 1962 (B): Meet your basic expenses with a little left over for extras  
 1963 (C): Just meet your basic expenses  
 1964 (D): Don’t even have enough to meet basic expenses  
 1965 (E): Refused

1966 **Participants:** [roughly 10,000] paid participants from a representative sample of the US populace.

1967

1968 **Participant grouping variables** (n=13): *CREGION*: U.S. region of residence, *AGE*: age bracket of  
 1969 the respondent, *SEX*: male or female, *EDUCATION*: highest level of education completed, *CITIZEN*:  
 1970 the respondent is (not) a citizen of the US, *MARITAL*: current marital status, *RELIG*: religious  
 1971 affiliation, *RELIGATTEND*: frequency of religious service attendance, *POLPARTY*: political party  
 1972 affiliation, *INCOME*: income bracket, *POLIDEOLOGY*: political ideology (e.g., liberal/conservative),  
 1973 *RACE*: racial identity.

1974 **Default System Prompt:**

1975

1976 You are from the United States.

1977

1978 **License:** No licensing information provided; Data is freely available with-  
 1979 out registration at <https://worksheets.codalab.org/worksheets/0x6fb693719477478aac73fc07db333f69>

1980

1981 **Publication:** Santurkar et al. (2023)

1982

1983 L.5 MORALMACHINECLASSIC

1984

1985 **Description:** This dataset contains three **moral decision-making scenarios**, which a large number  
 1986 of participants were asked to provide **subjective choices** for. The data was originally collected to  
 1987 study universals and variations in moral decision-making across the world.

1988 **Questions:** 3, with an average of 17,720 responses per question.

1989

1990 **Example question:**

1991

1992 A man in blue is standing by the railroad tracks when he notices an empty boxcar rolling out of  
 1993 control. It is moving so fast that anyone it hits will die. Ahead on the main track are five people.  
 1994 There is one person standing on a side track that doesn’t rejoin the main track. If the man in blue  
 1995 does nothing, the boxcar will hit the five people on the main track, but not the one person on the  
 1996 side track. If the man in blue flips a switch next to him, it will divert the boxcar to the side track  
 1997 where it will hit the one person, and not hit the five people on the main track. What should the man  
 in blue do?

1998 **Participants:** 19,720 volunteer participants (sociodemographics recorded) who chose to share their  
 1999 choices on the Moral Machine Classic web interface .

2000  
 2001 **Participant grouping variables** (n=6): *country*: respondent's country of residence, *gender*: gender of  
 2002 the respondent, *education*: level of education, *age\_group*: age bracket, *political\_group*: self-identified  
 2003 political orientation, *religious\_group*: self-identified religious affiliation.

2004 **Default System Prompt:**

2005  
 2006 The Moral Machine website (moralmachine.mit.edu) was designed to collect large-scale data on  
 2007 the moral acceptability of moral dilemmas. You are a user of the Moral Machine website.

2008 **License:** No licensing information provided.

2009 **Publication:** Awad et al. (2020)

2010 **L.6 CHAOSNLI**

2011 **Description:** This dataset contains **natural language inference scenarios** which participants were  
 2012 asked to provide **subjective judgments** on. The data was originally collected to study human  
 2013 disagreement on natural language inference scenarios.

2014 **Questions:** 4,645, with exactly 100 responses per question.

2015 **Example question:**

2016 Given a premise and a hypothesis, determine if the hypothesis is true (entailment), false  
 2017 (contradiction), or undetermined (neutral) based on the premise.

2018 Premise: Two young children in blue jerseys, one with the number 9 and one with the number 2  
 2019 are standing on wooden steps in a bathroom and washing their hands in a sink.

2020 Hypothesis: Two kids at a ballgame wash their hands.

2021 Choose the most appropriate relationship between the premise and hypothesis:

2022 (A): Entailment (the hypothesis must be true if the premise is true)  
 2023 (B): Contradiction (the hypothesis cannot be true if the premise is true)  
 2024 (C): Neutral (the hypothesis may or may not be true given the premise)

2025 **Participants:** 5,268 Amazon Mechanical Turk workers.

2026 **Participant grouping variables:** None.

2027 **Default System Prompt:**

2028 You are an Amazon Mechanical Turk worker.

2029 **License:** CC BY-NC 4.0

2030 **Publication:** Nie et al. (2020)

2031 **L.7 EUROPEAN SOCIAL SURVEY (ESS)**

2032 **Description:** This dataset contains three waves of **survey questions** that ask participants to provide  
 2033 **self-assessments** and **subjective judgments**. The data was originally collected to study attitudes and  
 2034 behaviors across the European populace. We use ESS wave 8-10.

2035 **Questions:** 237, with an average of 41,540 responses per task.

2036 **Example question:**

2037 Sometimes the government disagrees with what most people think is best for the country. Which  
 2038 one of the statements on this card describes what you think is best for democracy in general?

2052

2053

2054

2055

2056

Options:

- (A): Government should change its policies
- (B): Government should stick to its policies
- (C): It depends on the circumstances

2057

2058

**Participants:** Around 40,000 participants in total from European countries.

2059

2060

2061

2062

2063

2064

2065

**Participant grouping variables** (n=14): *cntry*: respondent's country of residence, *age\_group*: age bracket, *gndr*: gender of the respondent, *eisced*: level of education (ISCED classification), *household\_size\_group*: size of the household, *mnactic*: main activity status, *rlgdgr*: degree of religiosity, *lrscale*: self-placement on left-right political scale, *brncntr*: born in the country or abroad, *ctzcntr*: citizenship status, *domicil*: urban or rural living environment, *dscrgrp*: member of a group discriminated against, *uemp3m*: unemployed in the last 3 months, *maritalb*: marital status (married, single, separated, etc.)

2066

**Default System Prompt:**

2067

2068

The year is {survey year}.

2069

2070

**License:** CC BY-NC-SA 4.0

2071

**Publication:** European Social Survey European Research Infrastructure (ESS ERIC) (2024)

2072

2073

2074

## L.8 AFROBAROMETER

2075

**Description:** Afrobarometer is an annual public opinion survey conducted across more than 35 African countries. It collects data on citizens' perceptions of democracy, governance, the economy, and civil society, asking respondents for **self-assessments** and **subjective judgments**. We use the data from the 2023 wave of the survey, obtained from the afrobarometer.org website. We use Afrobarometer Round 9.

2081

**Questions:** 213, with an average of 52,900 responses per question.

2082

2083

Do you think that in five years' time this country will be more democratic than it is now, less democratic, or about the same?

2084

Options:

2088

- (A): Much less democratic
- (B): Somewhat less democratic
- (C): About the same
- (D): Somewhat more democratic
- (E): Much more democratic
- (F): Refused
- (G): Don't know

2085

2086

2087

**Participants:** 1,200-2,400 per country, 39 countries

2097

2098

2099

2100

2101

**Participant grouping variables** (n=11): *country*: respondent's country, *gender*: male or female, *education*: education level, *age\_group*: age bracket, *religion*: stated religion, *urban\_rural*: area of living, *employment*: job situation, *bank\_account*: whether respondent has a bank account, *ethnic\_group*: respondent's ethnicity, *subjective\_income*: how often to go without cash income, *discuss\_politics*: how often to discuss politics,

2102

2103

**Default System Prompt:**

2104

2105

The year is {survey year}.

**License:** No explicit language forbidding redistribute.

2106      **Publication:** Afrobarometer (2023)  
 2107  
 2108

2109      L.9    OSPSYCHBIG5  
 2110

2111      **Description:** This dataset contains a collection of anonymized **self-assessments** from the Big Five  
 2112      Personality Test, designed to evaluate individuals across five core personality dimensions.

2113      **Questions:** 50, with an average of 19,632 responses per question.

2114      **Example question:**

2115      Indicate your level of agreement with the following statement:  
 2116      I am always prepared.

2117      Options:

- 2118      (A): Disagree
- 2119      (B): Slightly Disagree
- 2120      (C): Neutral
- 2121      (D): Slightly Agree
- 2122      (E): Agree

2123      **Participants:** 19,719 volunteer participants from all over the world, who chose to share their  
 2124      assessments on the dedicated Open-Source Psychometrics web interface.

2125      **Participant grouping variables (n=3):** **country\_name:** country of residence, **gender\_cat:** male,  
 2126      female, or other, **age\_group:** age bracket.

2127      **Default System Prompt:**

2128      openpsychometrics.org is a website that provides a collection of interactive personality tests with  
 2129      detailed results that can be taken for personal entertainment or to learn more about personality  
 2130      assessment. You are a user of openpsychometrics.org.

2131      **License:** Creative Commons.

2132      **Publication:** None.

2133      L.10    OSPSYCHMGKT  
 2134

2135      **Description:** This dataset contains anonymized **test results** from the Multifactor General Knowledge  
 2136      Test (MGKT), a psychometric instrument designed to assess general knowledge across multiple  
 2137      domains. Each of the original 32 questions presents 10 answer options, of which 5 are correct. For  
 2138      consistency with other datasets in our study, we expand each question into 5 separate binary-choice  
 2139      items, each asking whether a given option is correct.

2140      **Questions:** 320, with an average of 18,644 responses per question.

2141      **Example question:**

2142      Is “Emily Dickinson” an example of famous poets?  
 2143      Choose one:  
 2144      (A) Yes  
 2145      (B) No

2146      **Participants:** 19,218 volunteer participants from all over the world, who chose to share their  
 2147      assessments on the dedicated Open-Source Psychometrics web interface.

2148      **Participant grouping variables (n=4):** **country\_name:** country of residence, **gender\_cat:** male,  
 2149      female, or other, **age\_group:** age bracket, **engnat\_cat:** is (not) a native English speaker.

2160  
 2161 openpsychometrics.org is a website that provides a collection of interactive personality tests with  
 2162 detailed results that can be taken for personal entertainment or to learn more about personality  
 2163 assessment. You are a user of openpsychometrics.org.

2164 **License:** Creative Commons.

2165 **Publication:** None.

2168 **L.11 OSPSYCHMACH**

2170 **Description:** This dataset contains anonymized **self-assessments** from the MACH-IV test, a psy-  
 2171 chometric instrument assessing the extent to which individuals endorse the view that effectiveness  
 2172 and manipulation outweigh morality in social and political contexts, i.e., their endorsement of  
 2173 Machiavellianism.

2174 **Questions:** 20, with an average of 54,974 responses per question.

2176 **Example question:**

2177 Indicate your level of agreement with the following statement:  
 2178 Never tell anyone the real reason you did something unless it is useful to do so.

2180 Options:

- 2181 (A): Disagree
- 2182 (B): Slightly disagree
- 2183 (C): Neutral
- 2184 (D): Slightly agree
- 2185 (E): Agree

2187 **Participants:** 73,489 volunteer participants from all over the world, who chose to share their  
 2188 assessments on the dedicated Open-Source Psychometrics web interface.

2189 **Participant grouping variables** (n=18): **country\_name**: country of residence, **gender\_cat**: male,  
 2190 female, or other, **age\_group**: age bracket, **race\_cat**: respondent's race, **engnat\_cat**: is (not) a native  
 2191 English speaker, **hand\_cat**: right-, left-, or both-handed, **education\_cat**: level of education, **urban\_cat**:  
 2192 type of urban area, **religion\_cat**: stated religion, **orientation\_cat**: sexual orientation, **voted\_cat**:  
 2193 did (not) vote at last elections, **married\_cat**: never, currently, or previously married, **familysize**:  
 2194 number of people belonging to the family, **TIPI\_E\_Group**: extraversion level based on TIPI score,  
 2195 **TIPI\_A\_Group**: agreeableness level based on TIPI score, **TIPI\_C\_Group**: conscientiousness level  
 2196 based on TIPI score, **TIPI\_ES\_Group**: emotional stability level based on TIPI score, **TIPI\_O\_Group**:  
 2197 openness-to-experience level based on TIPI score.

2198 openpsychometrics.org is a website that provides a collection of interactive personality tests with  
 2199 detailed results that can be taken for personal entertainment or to learn more about personality  
 2200 assessment. You are a user of openpsychometrics.org.

2202 **License:** Creative Commons.

2204 **Publication:** None.

2206 **L.12 OSPSYCHRWAS**

2208 **Description:** This dataset contains anonymized **self-assessments** from the Right-Wing Authoritarian-  
 2209 ism Scale (RWAS), a psychometric instrument assessing authoritarian tendencies such as submission  
 2210 to authority, aggression toward outgroups, and adherence to conventional norms.

2212 **Questions:** 22, with an average of 6,918 responses per question.

2213 **Example question:**

2214  
 2215 Please rate your agreement with the following statement on a scale from (A) Very Strongly  
 2216 Disagree to (I) Very Strongly Agree.  
 2217

2218 Statement: The established authorities generally turn out to be right about things, while the  
 2219 radicals and protestors are usually just "loud mouths" showing off their ignorance.  
 2220

Options:

- (A): Very Strongly Disagree
- (B): Strongly Disagree
- (C): Moderately Disagree
- (D): Slightly Disagree
- (E): Neutral
- (F): Slightly Agree
- (G): Moderately Agree
- (H): Strongly Agree
- (I): Very Strongly Agree

2230 **Participants:** 9,881 volunteer participants from all over the world, who chose to share their assess-  
 2231 ments on the dedicated Open-Source Psychometrics web interface.  
 2232

2233 **Participant grouping variables** (n=18): *age\_group*: age bracket, *gender\_cat*: male or female or  
 2234 other, *race\_cat*: respondent's race, *engnat\_cat*: is (not) English native, *hand\_cat*: right/left/both-  
 2235 handed, *education\_cat*: level of education, *urban\_cat*: type of urban area, *religion\_cat*: stated  
 2236 religion, *orientation\_cat*: sexual orientation, *voted*: did (not) vote at last elections, *married*:  
 2237 never/currently/Previously, *familysize*: number of people belonging to the family, *TIPI\_E\_Group*: ex-  
 2238 traversion level based on TIPI score, *TIPI\_A\_Group*: agreeableness level based on TIPI score,  
 2239 *TIPI\_C\_Group*: conscientiousness level based on TIPI score, *TIPI\_ES\_Group*: emotional sta-  
 2240 bility level based on TIPI score, *TIPI\_O\_Group*: openness-to-experience level based on TIPI  
 2241 score. *household\_income*: income sufficiency, *work\_status*: job situation, *religion*: stated religion,  
 2242 *nr\_of\_persons\_in\_household*: 1-7+, *marital\_status* respondent's legal relationship status, *domicil*:  
 2243 type of urban area,

2244 openpsychometrics.org is a website that provides a collection of interactive personality tests with  
 2245 detailed results that can be taken for personal entertainment or to learn more about personality  
 2246 assessment. You are a user of openpsychometrics.org.

2247 **License:** Creative Commons.  
 2248

2249 **Publication:** None.  
 2250

### 2251 L.13 INTERNATIONAL SOCIAL SURVEY PROGRAMME (ISSP)

2252 **Description:** The International Social Survey Programme (ISSP) is a **cross-national** collaborative  
 2253 programme conducting **annual surveys** on diverse **topics relevant to social sciences** since 1984. Of  
 2254 all 37 surveys, here we include only the five most recent surveys, which were collected in the years  
 2255 2017 to 2021.  
 2256

2257 **Questions:** 1,688, with an average of 7,074 responses per question.  
 2258

2259 **Participants:** 1,000 - 1,500 per country per wave  
 2260

2261 **Participant grouping variables** (n=11): *country*: respondent's country, *age*: age bracket, *gender*:  
 2262 male or female, *years\_of\_education*: 1-10+, *household\_income*: income sufficiency, *work\_status*: job  
 2263 situation, *religion*: stated religion, *nr\_of\_persons\_in\_household*: 1-7+, *marital\_status* respondent's  
 2264 legal relationship status, *domicil*: type of urban area, *topbot*: self-assessed social class  
 2265

2266 **Default System Prompt:**  
 2267

2268 The timeframe is {survey timeframe}.

2269 **License:** "Data and documents are released for academic research and teaching."

2268 **Publication:** see wave-specific references below.  
2269

2270 **L.13.1 ISSP 2017 SOCIAL NETWORKS AND SOCIAL RESOURCES**  
2271

2272 **Example question:**  
2273

2274 This section is about who you would turn to for help in different situations, if you needed it.  
2275

2276 For each of the following situations, please tick one box to say who you would turn to  
2277 first. If there are several people you are equally likely to turn to, please tick the box for the one  
2278 you feel closest to.

2279 Who would you turn to first to help you around your home if you were sick and had to  
2280 stay in bed for a few days?

2281 Options:  
2282

2283 (A): Close family member  
2284 (B): More distant family member  
2285 (C): Close friend  
2286 (D): Neighbour  
2287 (E): Someone I work with  
2288 (F): Someone else  
2289 (G): No one  
2290 (H): Can't choose

2291 **Publication:** ISSP Research Group (2019)  
2292

2293 **L.13.2 ISSP 2018 RELIGION IV**  
2294

2295 **Example question:**  
2296

2297 Please indicate which statement below comes closest to expressing what you believe about God.  
2298

2299 Options:  
2300

2301 (A): I don't believe in God  
2302 (B): Don't know whether there is a God and no way to find out  
2303 (C): Don't believe in a personal God, but in a Higher Power  
2304 (D): Find myself believing in God sometimes, but not at others  
2305 (E): While I have doubts, I feel that I do believe in God  
2306 (F): I know God really exists and have no doubts about it  
2307 (G): Don't know

2308 **Publication:** ISSP Research Group (2020)  
2309

2310 **L.13.3 ISSP 2019 SOCIAL INEQUALITY V**  
2311

2312 **Example question:**  
2313

2314 Looking at the list below, who do you think should have the greatest responsibility for reducing  
2315 differences in income between people with high incomes and people with low incomes?

2316 Options:  
2317

2318 (A): Can't choose  
2319 (B): Private companies  
2320 (C): Government  
2321 (D): Trade unions  
2322 (E): High-income individuals themselves

2322 (F): Low-income individuals themselves  
2323 (G): Income differences do not need to be reduced  
2324

2325 **Publication:** ISSP Research Group (2022)  
2326

2327 L.13.4 ISSP 2020 ENVIRONMENT IV  
2328

2329 **Example question:**

2330 In the last five years, have you ...  
2331

2332 Taken part in a protest or demonstration about an environmental issue?  
2333

2334 Options:  
2335

2336 (A): Yes, I have  
2337 (B): No, I have not

2338 **Publication:** ISSP Research Group (2023)  
2339

2340 L.13.5 ISSP 2021 HEALTH AND HEALTH CARE II  
2341

2342 **Example question:**

2343 During the past 12 months, how often, if at all, have you used the internet to look for information  
2344 on the following topics?  
2345

2346 Information related to anxiety, stress, or similar problems?  
2347

2348 Options:  
2349

2350 (A): Can't choose  
2351 (B): Never  
2352 (C): Seldom  
2353 (D): Sometimes  
2354 (E): Often  
2355 (F): Very often

2356 **Publication:** ISSP Research Group (2024)  
2357

2358 L.14 LATINOBARÓMETRO  
2359

2360 **Description:**

2361 LatinoBarómetro is an annual public opinion survey conducted across 18 Latin American countries.  
2362 It gathers data on the state of democracies, economies, and societies in the region, asking for **self-**  
2363 **assessments** and **subjective judgments**. We use the data from the 2023 wave of the survey, obtained  
2364 from the latinobarometro.org website.

2365 **Questions:** 155, with an average of 18,083 responses per question.  
2366

2367 **Example question:**

2368 Generally speaking, would you say you are satisfied with your life? Would you say you are...  
2369

2370 (A): Does not answer  
2371 (B): Do not know  
2372 (C): Very satisfied  
2373 (D): Quite satisfied  
2374 (E): Not very satisfied  
2375 (F): Not at all satisfied

2376  
 2377 **Participants:** In total, 19,205 interviews were applied in 17 countries. Samples of 1,000 representative  
 2378 cases of each country were applied to the five Central American countries and the Dominican  
 2379 Republic, while for the other countries representative samples had size 1,200.

2380 **Participant grouping variables (n=11):** *country*: respondent's country, *age\_group*: age bracket,  
 2381 *gender*: male or female, *highest\_education*: education level, *household\_income*: income suffi-  
 2382 ciency, *employment\_status*: job situation, *religiosity*: degree of religiosity, *religion*: stated religion,  
 2383 *political\_group*: government vs opposition, *citizenship*: citizen or not, *city\_size*: urban area size.

2384 **Default System Prompt:**

2385     The year is {survey year}.

2386 **License:** No explicit language forbidding redistribute.

2387 **Publication:** Latinobarómetro (2023)

2390 **L.15 GLOBALOPINIONQA**

2391 **Description:** This dataset contains survey questions that ask participants to provide **self-assessments**  
 2392 and **subjective judgments**, covering topics such as democracy, governance, international relations,  
 2393 and social values. The data was sourced from the World Values Survey and Pew Global Attitudes Sur-  
 2394 vey, and then repurposed to evaluate LLM alignment with the opinions of different sociodemographic  
 2395 groups.

2396 **Questions:** 2,556, grouping detail unknown

2397 **Example question:**

2400     Question: Do you strongly agree, agree, disagree, or strongly disagree with the following  
 2401 statement:  
 2402        "On the whole, men make better business executives than women do."

2403        (A) Agree strongly  
 2404        (B) Agree  
 2405        (C) Disagree  
 2406        (D) Strongly disagree  
 2407        (E) Don't know

2408 **Participants:** Pew Global Attitudes Survey: around 1,000 per country; World Values Survey: around  
 2409 1,200 per country

2410 **Participant grouping variables (n=1):** country

2411 **Default System Prompt:** None.

2412 **License:** CC BY-NC-SA 4.0

2413 **Publication:** Durmus et al. (2024)

2414 **L.16 DIVERSITY IN CONVERSATIONAL AI EVALUATION FOR SAFETY (DICES) - 990**

2415 **Description:** This dataset comprises 990 multi-turn adversarial conversations between humans and  
 2416 a dialogue model, accompanied by raters' **judgement** on the safety of last model response. This  
 2417 dataset captures variability in human perception, making it especially valuable for analyzing how  
 2418 demographic diversity influences safety assessments of conversational AI.

2419 **Questions:** 27,748 (990 turn annotated in multiple dimensions), with an average of 72 responses per  
 2420 question.

2421 **Example question:**

2430 Conversation:  
 2431 A: :)  
 2432 B: Are you going to be okay?  
 2433 A: Yes I am fine  
 2434 B: Oh, well that's good to hear. Glad it all worked out.  
 2435 A: Yes  
 2436  
 2437 Bot's response: You don't have to talk to me if you don't want to though.  
 2438  
 2439 Regarding the \*last chatbot response\*, considering \*all\* potential safety issues – including any  
 2440 type of harmful content (such as violence, unsafe advice, explicit material), any form of bias  
 2441 (such as hate speech, stereotypes, slurs), and any kind of misinformation (such as false claims,  
 2442 conspiracy theories) – do you perceive it to be unsafe overall for \*\*any\*\* reason?  
 2443  
 2444 (A): Yes  
 2445 (B): No  
 2446 (C): Unsure

2447 **Participants:** 173 raters balanced by gender and country (US, India)

2448 **Participant grouping variables** (n=4): *locale*: respondent's country (U.S./India), *age\_group*: age  
 2449 bracket, *gender*: male or female, *education*: education level.

2451 You are a crowd worker.

2453 **License:** CC BY-NC-SA 4.0

2454 **Publication:** Aroyo et al. (2023)

2456 L.17 NUMBERGAME

2458 **Description:** This dataset contains anonymized **judgments** from a numerical generalization task  
 2459 inspired by Tenenbaum's "number game" experiment. Responses reflect both rule-based (e.g., "even  
 2460 numbers") and similarity-based (e.g., "close to 50") generalization strategies, providing insight into  
 2461 the interplay of probabilistic reasoning and cognitive heuristics.

2462 **Questions:** 25,499, with an average of 10.15 responses per question.

2464 **Example question:**

2466 A program produces the following numbers: 63\_ 43.

2467 Is it likely that the program generates this number next: 24?

2469 (A): Yes

2470 (B): No

2471 **Participants:** 575 participants from the U.S.

2473 **Participant grouping variables** (n=4): *state*: respondent's state of residency in the U.S., *age\_group*:  
 2474 age bracket, *gender*: male or female, *education*: education level.

2475 You are an Amazon Mechanical Turk worker from the United States.

2476 **License:** CC0 1.0.

2479 **Publication:** Bigelow & Piantadosi (2016)

2481 L.18 CONSPIRACYCORR

2482 **Description:** This dataset contains **judgments** measuring individual endorsement of 11 widely  
 2483 circulated conspiracy theory beliefs.

2484 **Questions:** 9, with an average of 26,416 responses per question.  
 2485  
 2486

2487 **Example question:**  
 2488

2489 Would you say the following statement is true or false?  
 2490

2491 Statement: The US Government knowingly helped to make the 9/11 terrorist attacks happen in  
 2492 America on 11 September, 2001  
 2493

2494 Options:  
 2495

2496 (A): Definitely true  
 2497 (B): Probably true  
 2498 (C): Probably false  
 2499 (D): Definitely false  
 2500 (E): Don't know

2501 **Participants:** 26,416 participants from 20 different countries.  
 2502

2503 **Participant grouping variables** (n=4): *Country*: country of origin, *Age\_Group*: age bracket of the  
 2504 respondent, *Gender*: gender of the respondent, *Gender*: highest level of education completed  
 2505

2506 The year is {survey year}.  
 2507

2508 **License:** CC0 1.0 Universal.  
 2509

2510 **Publication:** Enders et al. (2024)  
 2511

## 2512 L.19 MORALMACHINE

2513 **Description:** This dataset contains responses from the Moral Machine experiment, a large-scale  
 2514 online platform designed to explore moral **decision-making** in the context of autonomous vehicles.  
 2515 Participants were asked to make ethical choices in life-and-death traffic scenarios, revealing  
 2516 preferences about whom a self-driving car should save.

2517 **Questions:** 2,073, with an average of 4,601 responses per question.  
 2518

2519 **Example question:**  
 2520

2521 You will be presented with descriptions of a moral dilemma where an accident is imminent and  
 2522 you must choose between two possible outcomes (e.g., 'Stay Course' or 'Swerve'). Each outcome  
 2523 will result in different consequences. Which outcome do you choose?  
 2524

2525 Options:  
 2526

2527 (A): Stay, outcome: in this case, the self-driving car with sudden brake failure will continue ahead  
 2528 and drive through a pedestrian crossing ahead. This will result in the death of the pedestrians.  
 2529

2530 **Dead:**  
 2531

2532 \* 1 woman  
 2533

2534 \* 1 boy  
 2535 \* 1 girl

2536 (B): Swerve, outcome: in this case, the self-driving car with sudden brake failure will swerve and  
 2537 crash into a concrete barrier. This will result in the death of the passengers.  
 2538

2539 **Dead:**  
 2540

2541 \* 1 woman  
 2542

2543 \* 1 elderly man  
 2544

2545 \* 1 elderly woman

2546 **Participants:** 492,921 volunteer participants from all over the world, participating through The  
 2547 Moral Machine web interface.  
 2548

2549 **Participant grouping variables** (n=1): *UserCountry3*: participant country,  
 2550

2538  
2539  
2540

The Moral Machine website ([moralmachine.mit.edu](http://moralmachine.mit.edu)) was designed to collect large-scale data on the moral acceptability of moral dilemmas. You are a user of the Moral Machine website.

2541  
2542  
2543

**License:** No formal open license is declared. However, the authors explicitly state that the dataset may be used beyond replication to answer follow-up research questions.

2544

**Publication:** Awad et al. (2018)

2545

## L.20 TRUST IN SCIENCE AND SCIENCE-RELATED POPULISM (TISP)

2547

**Description:** This dataset includes **judgements** about individuals' perception of science, its role in society and politics, attitudes toward climate change, and science communication behaviors.

2550

**Questions:** 97, with an average of 69.234 responses per question.

2551

**Example question:**

2553  
2554

How concerned or not concerned are most scientists about people's wellbeing?

2555

Options:

2556

- (A): not concerned
- (B): somewhat not concerned
- (C): neither nor
- (D): somewhat concerned
- (E): very concerned

2561  
2562

**Participants:** 71,922 participants across 68 countries.

2563  
2564  
2565  
2566

**Participant grouping variables** (n=8): *country*: respondent's country, *gender*: male or female, *age\_group*: age bracket, *education*: education level, *political\_alignment*: political stance (e.g., conservative), *religion*: level of religious belief, *residence*: type of living area (e.g., urban, rural), *income\_group*: income bracket.

2567  
2568

The year is {survey year}.

2569  
2570

**License:** no explicit language forbidding redistribute.

2571  
2572

**Publication:** Mede et al. (2025)

2573  
2574

## M ADDITIONAL RELATED WORK

2575

**Benchmarks for LLM Evaluation** Comprehensive benchmarks have been instrumental in driving LLM advancement by providing standardized evaluation frameworks. General language understanding benchmarks such as GLUE (Wang et al., 2018) and MMLU (Hendrycks et al., 2021) have established foundational metrics for assessing natural language understanding and reasoning capabilities. As LLM applications have diversified, domain-specific benchmarks have emerged, including TruthfulQA (Lin et al., 2022) for factual accuracy, LegalBench (Guha et al., 2023) for legal reasoning, and Chatbot Arena (Chiang et al., 2024) for chat assistants. These specialized benchmarks have enabled more precise evaluation of LLMs' fitness for particular use cases and have guided domain-specific optimization.

2584  
2585  
2586  
2587  
2588  
2589  
2590  
2591

Most closely related to SIMBENCH are OpinionQA (Santurkar et al., 2023) and GlobalOpinionQA (Durmus et al., 2024), which evaluate how accurately LLMs represent viewpoints of specific demographic groups. However, these benchmarks are limited in scope: OpinionQA focuses exclusively on U.S. public opinion surveys, while GlobalOpinionQA extends this approach globally but remains constrained to survey data. In contrast, SIMBENCH represents a substantial advancement in simulation evaluation by: (1) incorporating a diverse collection of 20 distinct tasks spanning multiple domains beyond surveys, (2) conceptualizing simulation as a fundamental capability deserving systematic evaluation rather than merely a representation challenge, and (3) establishing a unified evaluation framework that enables consistent cross-domain and cross-model comparison of simulation fidelity.

**Distribution Elicitation Methodologies** Prior research has primarily relied on first token probabilities to obtain survey answers from LLMs (Santurkar et al., 2023; Dominguez-Olmedo et al., 2024; Tjuatja et al., 2024). Unlike typical language model applications that focus on the model’s most likely completion, group-level LLM simulations aim to obtain normalized probabilities across all answer options. Recent work has demonstrated that verbalized responses yield better results for this purpose (Tian et al., 2023; Meister et al., 2025). Nevertheless, calibration of LLM outputs remains an open challenge; while extensively studied for model answer confidence (Zhao et al., 2021; Jiang et al., 2021; Kapoor et al., 2024; Zhu et al., 2023) and hallucinations (Kalai & Vempala, 2024), these issues also apply to simulating population response distributions. While instruction tuning can enhance models’ ability to produce accurate verbalized outputs, it may simultaneously impair calibration of normalized answer option probabilities (Cruz et al., 2024).

**Simulation of Complex Human Behavior** Few recent works have investigated LLM capabilities for simulation of temporal changes in human behavior Lazaridou et al. (2021). Ahnert et al. (2024) propose temporal adapters for LLMs for longitudinal analysis. While promising, such approaches remain constrained by limited availability of high-quality longitudinal datasets that capture human behavior changes over time.

More complex simulation of human social dynamics has been explored through multi-agent frameworks. Park et al. (2024a) developed large-scale simulations with LLM-powered agents to model emergent social behaviors. These approaches extend beyond static response prediction, making reliable simulations of complex human behavior even more difficult.

## N LLM USAGE

In this work, LLMs and AI-powered coding assistants were utilized as assistive tools. For paper writing, LLMs were used to rephrase and refine drafted paragraphs to improve clarity and readability. The authors then performed manual edits to ensure the final text was accurate and aligned with our intended meaning. For the implementation, we used AI-powered code editors and assistants, specifically Cursor and GitHub Copilot. These tools aided in writing and debugging Python scripts for data analysis.

## O DETAILED DEMOGRAPHIC BREAKDOWN

Table 13 provides a detailed count of the number of simulation targets (samples) associated with specific countries or regions across the `SIMBENCHPOP` and `SIMBENCHGROUPED` splits.

Table 13: Distribution of samples by Country/Region. `SIMBENCHPOP` represents general population questions, while `SIMBENCHGROUPED` represents specific demographic conditioned queries.

Country/Region	SimBenchPop	SimBenchGrouped	Total
Albania	0	5	5
Angola	12	11	23
Argentina	37	51	88
Armenia	0	2	2
Australia	6	188	194
Austria	72	52	124
Bangladesh	0	20	20
Belarus	0	1	1
Belgium	90	39	129
Benin	39	10	49
Bolivia	120	33	153
Botswana	52	24	76
Brazil	72	104	176
Bulgaria	78	20	98

Continued on next page

Table 13 – continued from previous page

Country/Region	SimBenchPop	SimBenchGrouped	Total
Burkina Faso	45	11	56
Cabo Verde	14	7	21
Cameroon	90	21	111
Canada	0	203	203
Chile	66	51	117
China	9	38	47
Colombia	105	33	138
Costa Rica	35	37	72
Côte d'Ivoire	75	28	103
Croatia	12	27	39
Cyprus	10	8	18
Czech Republic	108	49	157
Democratic Republic of the Congo	0	3	3
Denmark	60	42	102
Dominican Republic	90	33	123
Ecuador	79	36	115
Egypt	0	16	16
El Salvador	36	25	61
Estonia	40	26	66
Eswatini	39	9	48
Ethiopia	80	26	106
Finland	64	54	118
France	149	132	281
Gabon	30	11	41
Gambia	42	14	56
Georgia	10	11	21
Germany	146	236	382
Ghana	50	20	70
Greece	0	28	28
Guatemala	88	33	121
Guinea	28	17	45
Honduras	57	31	88
Hong Kong	0	10	10
Hungary	92	53	145
Iceland	36	45	81
India	7	343	350
Indonesia	0	16	16
Iran	0	2	2
Iraq	0	1	1
Ireland	102	35	137
Israel	70	39	109
Italy	105	70	175
Japan	52	36	88
Jordan	0	10	10
Kazakhstan	0	12	12
Kenya	100	34	134
Kuwait	0	3	3
Latvia	0	2	2
Lebanon	0	12	12
Lesotho	28	16	44
Liberia	12	13	25
Lithuania	113	43	156
Madagascar	42	12	54
Malawi	48	8	56

Continued on next page

Table 13 – continued from previous page

Country/Region	SimBenchPop	SimBenchGrouped	Total
Malaysia	0	18	18
Mali	14	16	30
Mauritania	52	14	66
Mauritius	12	12	24
Mexico	65	63	128
Mongolia	0	1	1
Montenegro	20	5	25
Morocco	13	32	45
Mozambique	33	12	45
Myanmar	0	2	2
Namibia	45	10	55
Netherlands	35	41	76
New Zealand	18	37	55
Nicaragua	0	9	9
Niger	52	16	68
Nigeria	19	41	60
North Macedonia	0	10	10
Norway	25	31	56
Pakistan	0	15	15
Palestinian Territory	0	5	5
Panama	45	29	74
Paraguay	43	24	67
Peru	119	39	158
Philippines	37	70	107
Poland	111	61	172
Portugal	80	28	108
Puerto Rico	0	2	2
Republic of the Congo	30	11	41
Romania	0	12	12
Russia	159	85	244
São Tomé and Príncipe	26	8	34
Senegal	42	14	56
Serbia	20	9	29
Seychelles	14	19	33
Sierra Leone	45	9	54
Singapore	0	4	4
Slovakia	50	29	79
Slovenia	89	42	131
South Africa	54	53	107
South Korea	5	24	29
Spain	61	63	124
Sudan	13	5	18
Suriname	10	15	25
Sweden	71	52	123
Switzerland	132	38	170
Taiwan	26	21	47
Tajikistan	0	1	1
Tanzania	46	19	65
Thailand	48	20	68
Togo	30	11	41
Tunisia	11	18	29
Turkey	24	25	49
Türkiye	0	5	5
Uganda	48	22	70

Continued on next page

**Table 13 – continued from previous page**

<b>Country/Region</b>	<b>SimBenchPop</b>	<b>SimBenchGrouped</b>	<b>Total</b>
Ukraine	0	11	11
United Kingdom	74	263	337
United States	1028	1636	2664
Uruguay	54	42	96
Uzbekistan	0	2	2
Venezuela	147	45	192
Vietnam	0	8	8
Zambia	64	16	80
Zimbabwe	42	18	60
<b>Grand Total</b>	<b>6,009</b>	<b>6,343</b>	<b>12,352</b>

2754  
2755  
2756  
2757  
2758  
2759  
2760  
2761  
2762  
2763  
2764  
2765  
2766  
2767  
2768  
2769  
2770  
2771  
2772  
2773  
2774  
2775  
2776  
2777  
2778  
2779  
2780  
2781  
2782  
2783  
2784  
2785  
2786  
2787  
2788  
2789  
2790  
2791  
2792  
2793  
2794  
2795  
2796  
2797  
2798  
2799  
2800  
2801  
2802  
2803  
2804  
2805  
2806  
2807