# SimBench: Benchmarking the Ability of Large Language Models to Simulate Human Behaviors

**Tiancheng Hu**[1], **Joachim Baumann**[2,3], **Lorenzo Lupo**[3], **Dirk Hovy**[3], **Nigel Collier**[1], and **Paul Röttger**[3]

[1]University of Cambridge, [2]University of Zurich, [3]Bocconi University,

Data Code

## Abstract

Simulations of human behavior based on large language models (LLMs) have the potential to revolutionize the social and behavioral sciences, *if and only if* they faithfully reflect real human behaviors. Prior work across many disciplines has evaluated the simulation capabilities of specific LLMs in specific experimental settings, but often produced disparate results. To move towards a more robust understanding, we introduce SimBench, the first large-scale benchmark to evaluate how well LLMs can simulate group-level human behaviors across diverse settings and tasks. SimBench compiles 20 datasets in a unified format, measuring diverse types of behavior (e.g., decision-making vs. self-assessment) across hundreds of thousands of diverse participants from different parts of the world. Using SimBench, we can ask fundamental questions regarding when, how, and why LLM simulations succeed or fail. For example, we show that, while even the best LLMs today have limited simulation ability, there is a clear log-linear scaling relationship with model size, and a strong correlation between simulation and scientific reasoning abilities. We also show that base LLMs, on average, are better at simulating high-entropy response distributions, while the opposite holds for instruction-tuned LLMs. By making progress measurable, we hope that SimBench can accelerate the development of better LLM simulators in the future.

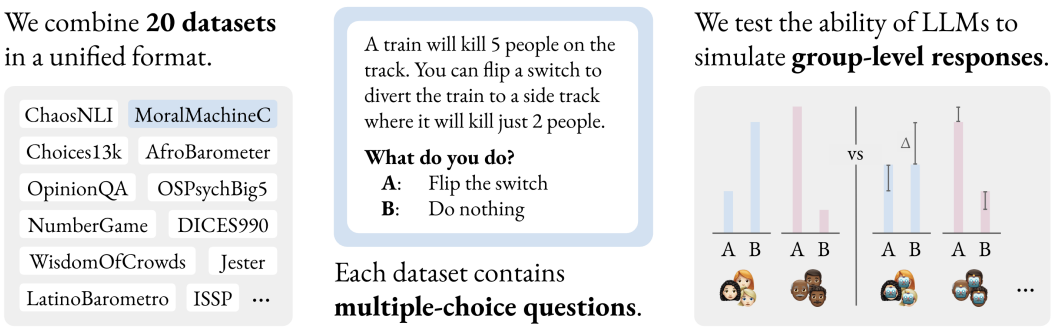


Figure 1: **SimBench** is the first-large scale benchmark to evaluate how well LLMs can simulate group-level human behavior across diverse simulation settings and tasks.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

# 1 Introduction

Large-scale human experiments and surveys have long been essential tools for informing public policy, commercial decisions, and academic research. Running experiments and surveys, however, is costly and time-consuming. Large language models (LLMs) can potentially address this challenge by simulating human behaviors quickly and at low cost, to complement or even substitute human studies. This prospect, alongside encouraging early evidence on the efficacy of LLMs as simulators [2, 6, 31], has motivated a large body of recent work across many disciplines investigating the ability of LLMs to simulate human behaviors [11, 12, 19, 46, 33, inter alia].

Most prior work, however, has been highly specific, evaluating the simulation ability of a narrow set of LLMs for a specific set of tasks, producing varied and sometimes even conflicting results (§5). Overall, the evidence on LLM simulation ability resembles an incomplete patchwork, making it difficult to draw any broader conclusions about when, how, and why LLM simulations fail, or how LLMs can be trained to be better simulators.

To remedy these issues and enable a more robust science of LLM simulation, we introduce SimBench, the first large-scale benchmark for evaluating the ability of LLMs to simulate human behaviors across diverse settings and tasks. SimBench combines 20 datasets in a unified and easily adaptable format, including popular datasets used in prior work as well as new datasets used for the first time (Figure 1). Together, these datasets measure the ability of LLMs to simulate several distinct types of human behavior (e.g., decision-making vs. self-assessment) across a diversity of human respondents (e.g., from different parts of the world). With SimBench, we take a first step towards answering six fundamental research questions about the simulation ability of LLMs:

**RQ1**: How well can current LLMs simulate human behaviors across diverse settings and tasks?

We test 24 state-of-the-art LLMs (§3), and show that even the best LLMs today struggle to faithfully simulate group-level human behaviors (§4.1). Predictions from the best-performing LLM, on average, are closer to a uniform response baseline than the true human response distribution.

**RQ2**: How do LLM characteristics such as model size affect LLM simulation ability?

We show that simulation ability grows log-linearly with model size (§4.2). We also find indicative evidence that increasing test-time compute does not meaningfully improve LLM simulations.

**RQ3**: How does task selection affect LLM simulation fidelity?

We find that simulation fidelity varies substantially across tasks, with even the best LLM simulators consistently performing worse than a uniform response baseline on several datasets (4.3).

**RQ4**: How does the degree of human response plurality affect LLM simulation fidelity?

We find that instruction-tuned LLMs tend to perform better on questions where humans give similar answers whereas base LLMs tend to perform better on questions where humans differ (§4.4).

**RQ5**: Are LLMs better at simulating responses from some groups than others?

We show that, on SimBench, LLMs struggle more with simulating specific demographic groups, especially those based on religion and ideology, compared to general populations (§4.5).

**RQ6**: To what extent does LLM simulation ability correlate with different model capabilities?

We find positive correlations with several popular capability benchmarks, including a particularly strong correlation with performance on scientific reasoning tasks (§4.6).

Progress in AI is only possible through rigorous evaluation, and large-scale benchmarks such as MMLU [29] have significantly contributed to improvements in LLM capabilities. We hope that

SimBench can play a similar role in accelerating the development of LLMs for simulating human behaviors. All of SimBench is permissively licensed and available on GitHub and Hugging Face.

# 2 Creating SimBench

## 2.1 Selecting Datasets for SimBench

To create SimBench, we conducted an open-ended search for suitable datasets in the social and behavioral sciences, guided by two main selection criteria: i) **large participant counts**, so that each dataset captures meaningful response distributions rather than the idiosyncratic behavior of few individuals; and ii) **permissive licensing** to freely redistribute each dataset as part of SimBench.

We generally opted for **datasets that have not been used to evaluate LLMs** in prior work, to increase the novelty and effectiveness of SimBench. However, to increase coverage and backward comparability, we also included datasets used in prior work (e.g., OpinionQA, ChaosNLI).

We also prioritized **datasets that provide participants' sociodemographic information** to evaluate the ability of LLMs to simulate responses from specific participant groups (see §2.3). Most survey datasets, for example, include this information. However, we also included three datasets that do not provide sociodemographic information (Jester, ChaosNLI, Choices13k) because they substantially increase the overall task diversity in SimBench.

Overall, SimBench includes 20 datasets, which we list in Appendix F, providing details on participants and example questions. Crucially, SimBench is fully modular by design, so that future work can easily add more datasets using the processing pipeline described in §2.2 below. In its release version, SimBench already meets two key criteria for comprehensive evaluation of LLM simulation ability:

1) **Task Diversity**: The 20 datasets in SimBench cover a wide range of different tasks regarding the human behavior they measure. SimBench includes **decision-making** questions (e.g., in Choices13k, MoralMachine), where participants are presented with a set of actions that concern themselves, and they have to select the action they would hypothetically take. SimBench also includes **self-assessment** questions (e.g., in OpinionQA, OSPsychBig5), where participants are presented with a set of descriptions or attributes, and they have to select the one that best describes themselves. Further, SimBench includes **judgment** questions (e.g., in ChaosNLI and Jester) where participants are presented with some external object and a choice of labels, and they have to select the label they think fits best. Lastly, SimBench includes **problem-solving** questions (e.g., in WisdomOfCrowds and OSPsychMGKT), where participants are presented with a set of answers to a factual question, and they have to select the answer they think is correct. Consequently, LLMs have to accurately simulate several distinct types of human behavior in order to perform well on SimBench.

2) **Participant Diversity**: The 20 datasets in SimBench capture a rich demographic landscape spanning at least 130 different countries across six continents. This global representation is a key strength of the benchmark. While five datasets include US-based crowdworkers, the international scope of SimBench is substantial: 3 datasets (e.g., LatinoBarometro, AfroBarometer) exclusively feature participants from regions outside the US, 4 datasets (e.g., GlobalOpinion, TISP) draw from multi-country samples across different continents, and 2 datasets collect responses from a global pool of internet users. Importantly, 8 out of the 20 datasets employ representative sampling techniques, enhancing the ecological validity of these constituent components. To perform well on SimBench, LLMs must therefore demonstrate the ability to accurately simulate the behavior of human participants across diverse cultural, linguistic, and socioeconomic backgrounds.[1]

## 2.2 Unifying SimBench Dataset Formats

**Question Selection & Format:** SimBench is a multiple-choice benchmark. From all 20 datasets, we therefore select only multiple-choice questions, and transform continuous scale questions into multiple-choice by splitting the scale into uniform bins. Where applicable, we collapse answer options to limit the maximum number of answering options to at most 26. In practice, questions rarely have more than 11 options. We exclude any questions with free-text answers and questions

[1] Note that, while some constituent datasets recruit representative samples, SimBench as a whole is not fully representative of any specific group of participants.

that are contingent on prior questions or with multi-turn interactions. For datasets with questions that are not originally in English, we use the English-language equivalents provided by the dataset creators. We do this to enable consistent evaluation, but we note that simulation ability may plausibly be correlated with prompt language, and encourage future work in this direction.

**Grouping Variables:** For each dataset, we record a brief description of the overall sampling population, the *default grouping*, in the form of a short prompt. For example, all participants in the WisdomOfCrowds dataset were US-based Amazon Mechanical Turk workers, so the default grouping prompt for this dataset is "You are an Amazon Mechanical Turk worker based in the United States.". Additionally, we select *grouping variables* for each dataset, corresponding to known participant sociodemographics, like age, gender, or race. The exact grouping variables and their values depend on what is available for each dataset. For a list of all grouping variables for each dataset, see Appendix F.

**Response Distributions:** We record the answers to each question in SimBench as group-level response distributions over the question's multiple-choice options. These distributions serve as the reference that we compare LLM predictions to. We create group-level response distributions by aggregating over the answers from all participants that belong to a given group. We set minimum grouping size thresholds for each dataset, filtering out groups with insufficient participants to form meaningful response distributions. Through this aggregation process, SimBench encompasses **10,930,271** unique question, grouping variable value pairs, each representing a distinct simulation target (see Table 3 for detailed counts). This approach enables robust evaluation of how accurately LLMs can simulate response patterns across diverse demographic groups and question types.

### 2.3 SimBench Splits

While the complete SimBench contains over 10 million potential test cases, for practical evaluation purposes we focus on two carefully curated splits that still provide comprehensive coverage of the simulation capabilities we aim to assess:

1) The **SimBenchPop** split covers all questions in all 20 datasets after processing as in §2.2. We combine each question with the dataset-specific default grouping prompt to create one unique test case, resulting in 7,167 test cases. We obtain the response distribution for each test case by aggregating all individual responses to that test case over all participants in that dataset. Conceptually, **SimBenchPop measures the ability of LLMs to simulate responses of broad and diverse human populations**.

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

## 3 Experimental Setup

**Tested Models:** To demonstrate the usefulness of SimBench and answer our six research questions (§1), we evaluate 24 state-of-the-art LLMs across 7 model families on SimBench. This includes both commercial and open-weight, base and instruction-tuned models, with model sizes ranging from 0.5B to 405B parameters. Table 1 shows the full list of models.

**Model Elicitation:** For each model, we collect predictions for the two main splits of SimBench (§2.3). To obtain model response distributions, we use one of two methods, depending on model type: 1) For base models, we directly extract **token probabilities** for each response option based on first-token logits. This is a natural way of eliciting a distribution out of an LLM, especially a base LLM. 2) For instruction-tuned models, we follow recent literature on LLM calibration and

[2]Ideally, we would also like to measure LLM simulation ability for intersectional groups that combine multiple characteristics (e.g., female + age 30-49). However, selecting on multiple characteristics substantially 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.

distribution prediction [63, 48] and use **verbalized distributions**, e.g., "Option A: 30%, Option B: 70%", elicited through prompting. For implementation details and prompt formats, see Appendix H.

**Evaluation Metric**: To measure LLM simulation ability, we derive the SimBench score $S$ from Total Variation Distance TVD, defined as:

$$S(P,Q) = 100\left(1 - \frac{TVD(P,Q)}{TVD(P,U)}\right) = 100\left(1 - \frac{\sum_i |P_i - Q_i|}{\sum_i |P_i - U_i|}\right) \quad (1)$$

where $P$ is the human ground truth distribution, $Q$ is the distribution predicted by the LLM that is being tested, and $U$ is a uniform distribution over all response options for a given question. Conceptually, $S$ therefore measures how much more accurate the predictions from an LLM are than predictions from a uniform baseline model, which assigns equal probability to all response options for a given question. In other words, $S$ quantifies the advantage of an LLM simulation over the simplest possible guess.

An $S$ score of 100 indicates perfect alignment between the LLM and the human ground truth distribution, while a score $\leq$0 indicates performance at or below the performance of a uniform baseline. We chose TVD as the basis for $S$ due to its symmetry, boundedness, and robustness to zero probabilities. For a comparison to alternative metrics, see Appendix D.

Table 1: **Overall simulation ability** as measured by SimBench score $S$ averaged across the two main splits of SimBench. Reasoning models are highlighted in *italics*. Models are sorted by score. Models below the dotted line perform worse than a uniform baseline.

| Model | Type | Release | $S$ (↑) |
|---|---|---|---|
| Claude-3.7-Sonnet | Instr. | Closed | 40.80 |
| *Claude-3.7-Sonnet-4000* | Instr. | Closed | 39.46 |
| GPT-4.1 | Instr. | Closed | 34.56 |
| *DeepSeek-R1* | Instr. | Open | 34.52 |
| DeepSeek-V3-0324 | Instr. | Open | 32.90 |
| *o4-mini-high* | Instr. | Closed | 28.99 |
| Llama-3.1-405B-Instruct | Instr. | Open | 28.41 |
| *o4-mini-low* | Instr. | Closed | 27.77 |
| Gemma-3-12B-IT | Instr. | Open | 18.63 |
| Gemma-3-27B-IT | Instr. | Open | 18.34 |
| Llama-3.1-70B-Instruct | Instr. | Open | 16.57 |
| Qwen2.5-72B | Base | Open | 13.35 |
| Qwen2.5-32B | Base | Open | 12.28 |
| Qwen2.5-14B | Base | Open | 11.93 |
| Qwen2.5-3B | Base | Open | 8.84 |
| Qwen2.5-7B | Base | Open | 8.76 |
| Gemma-3-12B-PT | Base | Open | 7.67 |
| Gemma-3-27B-PT | Base | Open | 5.54 |
| Qwen2.5-1.5B | Base | Open | 5.34 |
| Llama-3.1-8B-Instruct | Instr. | Open | -0.14 |
| Gemma-3-4B-PT | Base | Open | -0.73 |
| Gemma-3-4B-IT | Instr. | Open | -1.91 |
| Qwen2.5-0.5B | Base | Open | -2.99 |
| Gemma-3-1B-PT | Base | Open | -16.13 |

# 4 Results

## 4.1 RQ1: General Simulation Ability of LLMs

To evaluate the general simulation ability of LLMs, we measure their overall SimBench score $S$ averaged across the two main splits of SimBench (Table 1). We find that **even leading LLMs struggle to simulate group-level human behaviors with high accuracy**, as measured across the 20 datasets in SimBench. Claude-3.7-Sonnet is the best-performing model overall, but only achieves a score of 40.80 out of a maximum of 100 on SimBench. This score indicates that the response distributions predicted by Claude-3.7-Sonnet are, on average, closer to a uniform response distribution than to the true human response distribution. The distance from the true distribution is 19.7 percentage points, on average, as shown by the TVD listed in Table 5. The best-performing open-weight LLM is DeepSeek-R1, achieving a score of 34.52. The majority of the 24 models we test perform substantially worse still, scoring less than 20. Notably, five models we test score below 0, indicating that their predicted response distributions are, on average, even further away from the true human response distribution than a uniform response distribution. Overall, these results suggest that disparate results from prior work may combine into a somewhat disappointing picture, painting LLMs as far from reliable simulators when considering a diversity of tasks.

### 4.2 RQ2: Impact of LLM Characteristics on Simulation Ability

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

**1) Model Size** To evaluate the impact of model size on simulation ability, we plot SimBench Score $S$ against model parameter count for the four LLM families that we can test across multiple model sizes (Figure 2). Our results suggest that **there is a clear log-linear scaling law for LLM simulation ability**. Across all examined model families, an increase in parameter count generally corresponds to an increase in SimBench score $S$, indicating better alignment between predicted and human response distributions. Llama-3.1-Instruct in particular demonstrates nearly perfect log-linear scaling, with the largest Llama-3.1-405B-Instruct achieving a score of 28.41. Conversely, all models with low parameter counts ($\leq$10B) perform very poorly on SimBench, scoring at most 8.76 (Qwen2.5-7B). Overall, the clear positive scaling trends across model families suggest that, while simulation remains a challenging task for even the best models today, further model scaling may well lead to highly accurate LLM simulators in the future.

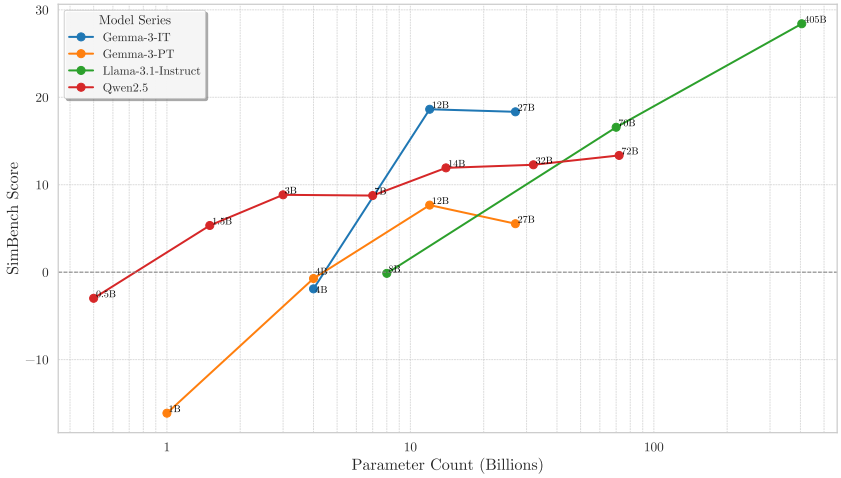


Figure 2: **Model parameter count vs. simulation ability**. We measure model size by parameter count and simulation ability by SimBench score $S$ averaged across the two main splits of SimBench.

**2) Test-Time Compute** To analyze the effects of increasing test-time compute on LLM simulation ability, we compare o4-mini-low vs. o4-mini-high, as well as Claude-3.7-Sonnet in its standard configuration vs. with a 4000-token thinking budget (Table 1). We are limited to these two comparisons due to budget constraints. Our results suggest that **there is no clear benefit to increasing test-time compute for LLM simulation ability**. However, this finding should only be interpreted as early, indicative evidence, and we hope that SimBench can enable further work in this direction.

### 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). Therefore, we break down simulation fidelity by dataset, showing results for the five LLMs we previously identified as the best simulators in Figure 3. We find that **simulation fidelity varies substantially across tasks**, with even the best LLM simulators performing worse than a uniform response baseline on several datasets, as indicated by negative SimBench scores (e.g., on Jester, OSPsychMach, and MoralMachine). Generally, the different LLMs exhibit similar performance patterns, with one notable exception being GPT-4.1's exceptionally high score of 61.9 on OSPsychRWAS.

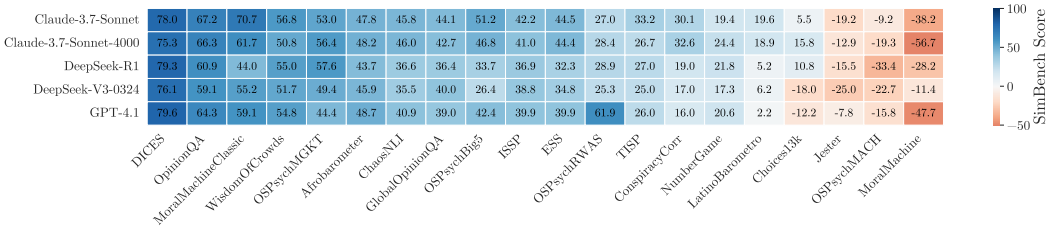


Figure 3: **Simulation fidelity by dataset** as measured by SimBench score $S$ for each of the 20 datasets in SimBenchPop. We show results for the top five models based on results in Table 1.

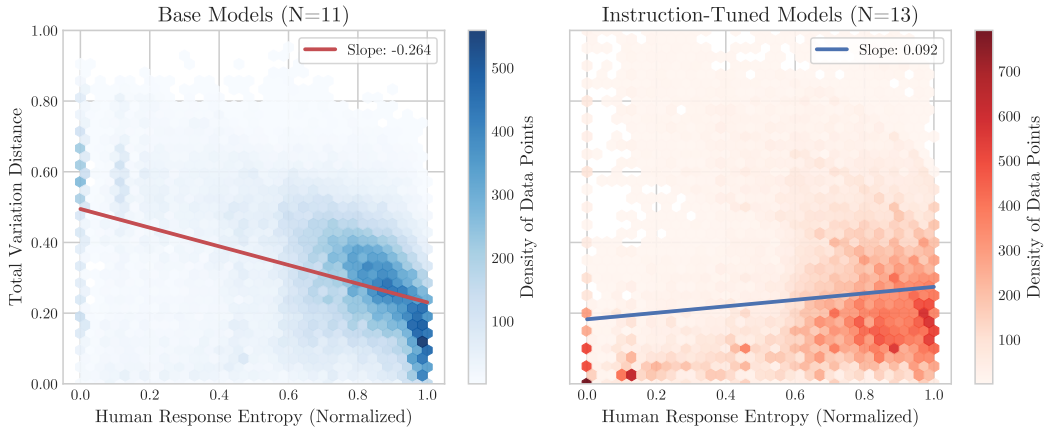


Figure 4: **Response plurality vs. simulation fidelity** for base and instruction-tuned models on all questions in SimBenchPop. We measure response plurality by normalised entropy of the human response distribution and simulation fidelity by total variation distance at the question level.

### 4.4 RQ4: Impact of Response Plurality on Simulation Fidelity

Human participants give very similar responses to some questions while giving very different responses to others. Faithful simulation requires models to perform well in either scenario. We operationalise the level of response plurality by measuring the normalised entropy of the human response distribution at the question level. We then plot this entropy for all questions in SimBenchPop against total variation distance (TVD, see §3), which measures the difference in predicted and reference distribution at a question level (Figure 4). Prior work has found that instruction-tuning encourages models to produce more confident, less ambiguous outputs, resulting in low-entropy token distributions [13, 63, 48, 16]. Therefore, we differentiate between base and instruction-tuned models for this analysis. We find that **base models generally perform better on questions where human participants tend to give different answers, whereas the inverse holds for instruction-tuned models**. This finding is supported by our regression analysis in Appendix 6, which confirms the statistical significance of this effect. Therefore, while instruction-tuned models tend to outperform base models in terms of overall score on SimBench (Table 1), our results here suggest that instruction-tuning also worsens simulation ability for at least a subset of high-plurality questions.

### 4.5 RQ5: Simulation Ability Across Participant Groups

Many applications require simulating responses from specific demographic groups rather than general populations. Using SimBenchGrouped, we evaluate how LLM simulation ability changes when conditioned on specific demographic attributes.

We measure this change as $\Delta S = S_{grouped} - S_{ungrouped}$, where $S_{ungrouped}$ is the SimBench score for simulating the general population and $S_{grouped}$ is the score when simulating a specific demographic group on the same question. A negative $\Delta S$ indicates that the model's simulation ability relative to the uniform baseline decreases when asked to simulate specific demographic groups.

Importantly, for SimBenchGrouped, we specifically selected questions where human response distributions showed the highest variance across demographic groups (see §2.3). The observed degradation in simulation performance therefore likely represents an upper bound on the challenges LLMs face when simulating specific demographic groups. Our results in Table 2 show that **LLMs struggle more with simulating specific demographic groups compared to general populations**. All evaluated models show negative mean $\Delta S$ values, with degradation ranging from -1.27 for DeepSeek-V3-0324 to -4.61 for Claude-3.7-Sonnet-4000.

Table 2: **Ungrouped vs. grouped** simulation performance $\Delta S$.

| **Models** | |
|---|---|
| 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 |

| **Demographics** | |
|---|---|
| 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 |

The performance degradation varies substantially by demographic category. Models struggle most when simulating groups defined by religious attributes, with conditioning on 'Religiosity/Practice' causing the largest decrease in simulation accuracy ($\Delta S = -9.91$), followed by 'Political Affiliation/Ideology' ($\Delta S = -4.97$) and 'Religion (Affiliation)' ($\Delta S = -4.83$). In contrast, models maintain relatively better performance when simulating groups defined by 'Gender' ($\Delta S = -1.24$) and 'Age' ($\Delta S = -1.50$).

While these findings may not fully generalize to cases where demographic differences are less pronounced, they highlight potential limitations in how current LLMs capture the nuanced response patterns of specific demographic groups. We argue that such challenging benchmarks are crucial for identifying areas where improvements are most needed, particularly for applications that aim to model the behaviors of specific subpopulations.

### 4.6 RQ6: Simulation Ability vs. General Capabilities

Finally, we analyze the relationship between LLM simulation ability and more general model capabilities by correlating performance on SimBench with popular LLM capability benchmarks (Figure 5). Specifically, we compare SimBench scores to performance on GPQA Diamond [59] and OTIS AIME [24], based on scores reported in the Epoch AI Benchmarking Hub [23], which we are able to retrieve for 8 of the LLMs we test. We also compare to Chatbot Arena ELO scores [15], retrieved for the same 8 models on May 14th, 2025.

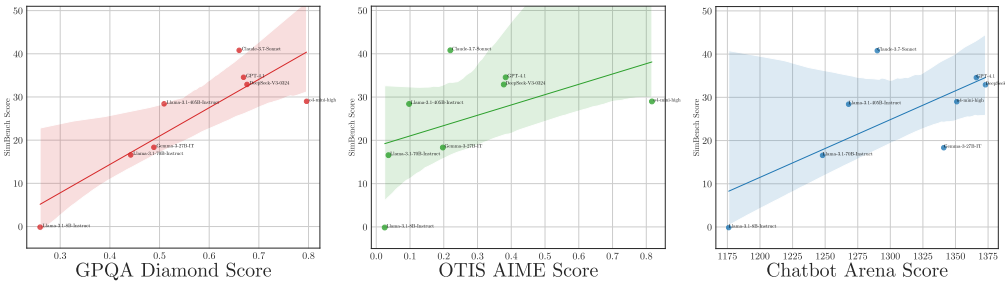


Figure 5: **General model capabilities vs. simulation ability**, as measured by popular benchmark scores compared to SimBench score $S$ averaged across the two main splits in SimBench.

We find that **simulation ability is positively correlated with general model capabilities**. This matches our earlier finding on the benefits of model scaling (§4.2). However, the strength of the correlation varies across capability benchmarks. Most notably, the very strong correlation with GPQA suggests that there may be substantial symbiotic effects between scientific reasoning and simulation for social and behavioral science tasks of the kind included in SimBench. By comparison, the weaker correlation with Chatbot Arena scores suggests optimising LLMs for general helpfulness and user satisfaction does not necessarily make them better simulators.

# 5 Related Work

**Human Behavior Simulation with LLMs** LLMs as human behavior simulators have attracted significant interdisciplinary attention. Researchers have evaluated their efficacy across political science [6, 12, 19], psychology [2, 11, 46, 30], economics [31, 2], and computer science applications [32, 20, 33, 55]. Evidence regarding LLMs' simulation fidelity remains mixed, with some studies reporting promising results [6] while others identify critical limitations, including homogenized group representations [14, 65] and deterministic rather than distributional predictions [57].

Existing work has predominantly focused on individual-level simulation with minimal demographic conditioning, typically evaluating only one or two models in narrowly defined contexts. SimBench addresses these limitations by providing a comprehensive benchmark for group-level simulation across diverse domains with systematic demographic conditioning and standardized metrics. The benchmark's distributional evaluation framework (using Total Variation distance) captures how accurately models represent the full spectrum of human response variation—an approach advocated by researchers in both simulation [4] and general LLM evaluation [69]. For broader context on this emerging field, we refer readers to recent comprehensive surveys [42, 52, 4].

**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 [66] and MMLU [29] have established foundational metrics for assessing natural language understanding and reasoning capabilities. As LLM applications have diversified, domain-specific benchmarks have emerged, including TruthfulQA [45] for factual accuracy, LegalBench [27] for legal reasoning, and Chatbot Arena [15] for chat assistants. These specialized benchmarks have enabled more precise evaluation of LLMs' fitness for particular use cases and have guided domain-specific optimization.

Most closely related to SimBench are OpinionQA [60] and GlobalOpinionQA [21], 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.

Appendix G continues our discussion of related work.

# 6 Conclusion

LLM simulations of human behavior have the potential to create immense benefits for society by helping shape effective policy, guiding industrial decisions, and informing academic research. To fulfill this potential, however, LLM simulations must be sufficiently faithful in representing real human behaviors across diverse settings and tasks. Prior work evaluating LLM simulation fidelity has taken a predominantly narrow approach, producing an incomplete patchwork of evidence.

To change this, we introduced SimBench, the first large-scale benchmark for evaluating group-level LLM simulation ability. We described the dataset selection and processing steps that resulted in 20 datasets with a unified format, measuring diverse types of human behavior (e.g., decision-making vs. self-assessment) across hundreds of thousands of diverse participants from different parts of the world. Using SimBench, we took a first step toward answering fundamental questions regarding when, how, and why LLM simulations succeed or fail. For example, we demonstrated that while even the best LLMs today have limited simulation ability, there is a clear log-linear scaling relationship with model size and a strong correlation between simulation and scientific reasoning abilities.

Significant progress remains to be made in developing LLMs as better simulators of human behavior. We hope that SimBench can provide an open foundation for future efforts in this direction, ultimately benefiting society as a whole.

# References


[1] Afrobarometer. Afrobarometer data, all countries (39), round 9, 2023. `http://www.afrobarometer.org`, 2023. Accessed: March 2025.

[2] G. V. Aher, R. I. Arriaga, and A. T. Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR, 23–29 Jul 2023.

[3] G. Ahnert, M. Pellert, D. Garcia, and M. Strohmaier. Extracting affect aggregates from longitudinal social media data with temporal adapters for large language models. *arXiv preprint arXiv:2409.17990*, 2024.

[4] J. R. Anthis, R. Liu, S. M. Richardson, A. C. Kozlowski, B. Koch, J. Evans, E. Brynjolfsson, and M. Bernstein. Llm social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*, 2025.

[5] Anthropic. Claude 3.7 sonnet and claude code, Feb. 2025. Accessed: 2025-05-14.

[6] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[7] L. Aroyo, A. S. Taylor, M. Díaz, C. M. Homan, A. Parrish, G. Serapio-García, V. Prabhakaran, and D. Wang. Dices dataset: diversity in conversational ai evaluation for safety. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[8] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

[9] E. Awad, S. Dsouza, A. Shariff, I. Rahwan, and J.-F. Bonnefon. Universals 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.

[10] E. Bigelow and S. T. Piantadosi. A large dataset of generalization patterns in the number game. *Journal of Open Psychology Data*, 4(1):e4–e4, 2016.

[11] M. Binz, E. Akata, M. Bethge, F. Brändle, F. Callaway, J. Coda-Forno, P. Dayan, C. Demircan, M. K. Eckstein, N. Éltető, et al. Centaur: a foundation model of human cognition. *arXiv preprint arXiv:2410.20268*, 2024.

[12] J. Bisbee, J. D. Clinton, C. Dorff, B. Kenkel, and J. M. Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416, 2024.

[13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[14] M. Cheng, T. Piccardi, and D. Yang. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875, Singapore, Dec. 2023. Association for Computational Linguistics.

[15] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. I. Jordan, J. E. Gonzalez, and I. 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.

[16] A. F. Cruz, M. Hardt, and C. Mendler-Dünner. Evaluating language models as risk scores. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 97378–97407. Curran Associates, Inc., 2024.

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

[18] T. Dettmers, M. Lewis, Y. Belkada, and L. 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.

[19] R. Dominguez-Olmedo, M. Hardt, and C. 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, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 45850–45878. Curran Associates, Inc., 2024.

[20] Y. R. Dong, T. Hu, and N. Collier. Can LLM be a personalized judge? In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics.

[21] E. Durmus, K. Nguyen, T. Liao, N. Schiefer, A. Askell, A. Bakhtin, C. Chen, Z. Hatfield-Dodds, D. Hernandez, N. Joseph, L. Lovitt, S. McCandlish, O. Sikder, A. Tamkin, J. Thamkul, J. Kaplan, J. Clark, and D. Ganguli. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*, 2024.

[22] A. Enders, C. Klofstad, A. Diekman, H. Drochon, J. Rogers de Waal, S. Littrell, K. Premaratne, D. Verdear, S. Wuchty, and J. Uscinski. The sociodemographic correlates of conspiracism. *Scientific reports*, 14(1):14184, 2024.

[23] Epoch AI. "ai benchmarking hub", 11 2024. https://epoch.ai/data/ai-benchmarking-dashboard.

[24] EpochAI. Otis mock aime 24-25. `https://huggingface.co/datasets/EpochAI/otis-mock-aime-24-25`, 2024. Accessed: 2025-05-11.

[25] European Social Survey European Research Infrastructure (ESS ERIC). ESS11 - Integrated File, Edition 2.0 [Data set], 2024.

[26] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *information retrieval*, 4:133–151, 2001.

[27] N. Guha, J. Nyarko, D. Ho, C. Ré, A. Chilton, A. K, A. Chohlas-Wood, A. Peters, B. Waldon, D. Rockmore, D. Zambrano, D. Talisman, E. Hoque, F. Surani, F. Fagan, G. Sarfaty, G. Dickinson, H. Porat, J. Hegland, J. Wu, J. Nudell, J. Niklaus, J. Nay, J. Choi, K. Tobia, M. Hagan, M. Ma, M. Livermore, N. Rasumov-Rahe, N. Holzenberger, N. Kolt, P. Henderson, S. Rehaag, S. Goel, S. Gao, S. Williams, S. Gandhi, T. Zur, V. Iyer, and Z. Li. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023.

[28] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[29] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

[30] L. Hewitt, A. Ashokkumar, I. Ghezae, and R. Willer. Predicting results of social science experiments using large language models, August 2024.

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

[32] T. Hu and N. Collier. Quantifying the persona effect in LLM simulations. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.

[33] T. Hu and N. Collier. inews: A multimodal dataset for modeling personalized affective responses to news. *arXiv preprint arXiv:2503.03335*, 2025.

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

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

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

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

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

[39] Z. Jiang, J. Araki, H. Ding, and G. Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

[40] A. T. Kalai and S. S. Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, STOC 2024, page 160–171, New York, NY, USA, 2024. Association for Computing Machinery.

[41] S. Kapoor, N. Gruver, M. Roberts, A. Pal, S. Dooley, M. Goldblum, and A. Wilson. Calibration-tuning: Teaching large language models to know what they don't know. In R. Vázquez, H. Celikkanat, D. Ulmer, J. Tiedemann, S. Swayamdipta, W. Aziz, B. Plank, J. Baan, and M.-C. de Marneffe, editors, *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 1–14, St Julians, Malta, Mar. 2024. Association for Computational Linguistics.

[42] A. C. Kozlowski and J. Evans. Simulating subjects: The promise and peril of ai stand-ins for social agents and interactions, September 2024. Preprint.

[43] Latinobarómetro. Latinobarómetro 2023. `http://www.latinobarometro.org`, 2023. Accessed: March 2025.

[44] A. Lazaridou, A. Kuncoro, E. Gribovskaya, D. Agrawal, A. Liska, T. Terzi, M. Gimenez, C. de Masson d'Autume, T. Kocisky, S. Ruder, D. Yogatama, K. Cao, S. Young, and P. Blunsom. Mind the gap: Assessing temporal generalization in neural language models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29348–29363. Curran Associates, Inc., 2021.

[45] S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[46] B. S. Manning, K. Zhu, and J. J. Horton. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research, 2024.

[47] N. G. Mede, V. Cologna, S. Berger, J. Besley, C. Brick, M. Joubert, E. W. Maibach, S. Mihelj, N. Oreskes, M. S. Schäfer, et al. Perceptions of science, science communication, and climate change attitudes in 68 countries–the tisp dataset. *Scientific data*, 12(1):114, 2025.

[48] N. Meister, C. Guestrin, and T. Hashimoto. Benchmarking distributional alignment of large language models. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 24–49, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics.

[49] Meta AI. Introducing Llama 3.1: Our most capable models to date, 2024. Accessed: 2025-05-14.

[50] Y. Nie, X. Zhou, and M. Bansal. What can we learn from collective human opinions on natural language inference data? In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online, Nov. 2020. Association for Computational Linguistics.

[51] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[52] A. Olteanu, S. Barocas, S. L. Blodgett, L. Egede, A. DeVrio, and M. Cheng. Ai automatons: Ai systems intended to imitate humans. *arXiv preprint arXiv:2503.02250*, 2025.

[53] OpenAI. Introducing GPT-4.1 in the API, 2025. Accessed: 2025-05-14.

[54] OpenAI. Introducing openai o3 and o4-mini, Apr. 2025. Accessed: 2025-05-14.

[55] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery.

[56] J. S. Park, C. Q. Zou, A. Shaw, B. M. Hill, C. Cai, M. R. Morris, R. Willer, P. Liang, and M. S. Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024.

[57] P. S. Park, P. Schoenegger, and C. Zhu. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 56:5754–5770, 2024.

[58] J. C. Peterson, D. D. Bourgin, M. Agrawal, D. Reichman, and T. L. Griffiths. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214, 2021.

[59] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

[60] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

[61] C. Simoiu, C. Sumanth, A. Mysore, and S. Goel. Studying the "wisdom of crowds" at scale. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 171–179, 2019.

[62] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[63] K. Tian, E. Mitchell, A. Zhou, A. Sharma, R. Rafailov, H. Yao, C. Finn, and C. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore, Dec. 2023. Association for Computational Linguistics.

[64] L. Tjuatja, V. Chen, T. Wu, A. Talwalkwar, and G. Neubig. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 09 2024.

[65] A. Wang, J. Morgenstern, and J. P. Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pages 1–12, 2025.

[66] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In T. Linzen, G. Chrupała, and A. Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics.

[67] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. In Q. Liu and D. Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics.

[68] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[69] L. Ying, K. M. Collins, L. Wong, I. Sucholutsky, R. Liu, A. Weller, T. Shu, T. L. Griffiths, and J. B. Tenenbaum. On benchmarking human-like intelligence in machines. *arXiv preprint arXiv:2502.20502*, 2025.

[70] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate before use: Improving few-shot performance of language models. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR, 18–24 Jul 2021.

[71] C. Zhu, B. Xu, Q. Wang, Y. Zhang, and Z. Mao. On the calibration of large language models and alignment. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9778–9795, Singapore, Dec. 2023. Association for Computational Linguistics.

# A Limitations

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

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

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

**Training Data Overlap** Without complete transparency into model training corpora, we cannot definitively rule out the possibility that some test items appeared during training. However, several factors mitigate concerns about data contamination affecting our results. First, SimBench evaluates simulation at the group distribution level rather than individual response prediction, making memorization of specific survey responses less impactful. Second, many of our datasets primarily exist as aggregated statistics in published research rather than as widely available raw data. Finally, the consistent scaling patterns we observe across diverse datasets suggest genuine simulation capabilities rather than artifacts of training data overlap. Nevertheless, we acknowledge that data contamination remains a fundamental challenge in LLM evaluation, and future work should develop more robust methods to detect and quantify its impact. We include this consideration for completeness while believing it unlikely to significantly impact our current findings.

# B Ethical Considerations

SimBench's primary purpose is to benchmark LLMs' ability to simulate human behavior. While advancements in LLM simulation capabilities can support helpful applications such as pre-testing policies, these do not come without risks of misrepresentation and dual use.

First and foremost, due to the observed limited simulation ability of state-of-the-art LLMs, we caution against relying on LLM-powered simulations of human behavior for tasks where downstream harm is possible. Even as models improve, substituting algorithmic approximations for authentic human participation carries the risk of disadvantaging under-represented/marginalized communities by removing their opportunities to directly shape decisions that affect them. Furthermore, while benchmarks like SimBench help measure simulation capabilities, we must be careful not to mistake increasing benchmark performance for genuine understanding of complex human behavior.

While SimBench includes diverse demographic groups, it can not adequately support simulations of intersectional identities due to sample size limitations. By conditioning on one demographic variable at a time, we cannot systematically assess how well models handle the rich overlap of identities (e.g., "older Latinx women," "young Black men"). Small intersectional group sizes make it difficult to combine multiple characteristics simultaneously due to increasing sampling noise in response distributions. Yet intersectional simulation is precisely where societal biases and model

limitations often emerge, making this an important direction for future work. Additionally, the conditional prompting approach we use conceptualizes simplistic human populations and may thus fail to appropriately account for nuances of individual behavior.

Nevertheless, we believe SimBench is an important step toward making LLM simulation progress measurable and raising awareness of state-of-the-art model blind spots. Together, we hope this will ultimately create accountability for models deployed in socially sensitive contexts.

# C SimBenchPop and SimBenchGrouped Sampling Details

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

**SimBenchPop** measures LLMs' ability to simulate responses of broad, diverse human populations. We include all questions from all 20 datasets in SimBench, combining each question with its dataset-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 keeping the benchmark manageable. For each test case, we aggregate individual responses across all participants in the dataset to create population-level response distributions. This approach creates a benchmark that represents population-level responses across diverse domains while maintaining a reasonable size of 7,167 test cases.

For **SimBenchGrouped**, we focus only on five large-scale survey datasets with rich demographic information and sufficient sample sizes: OpinionQA, ESS, Afrobarometer, ISSP, and LatinoBarometro. Our sampling approach prioritizes questions showing meaningful demographic variation. For each dataset, we identify available grouping variables (e.g., age, gender, country) with sufficient group sizes to form meaningful response distributions. We calculate the variance of responses across demographic groups for each question and rank questions by their variance scores, prioritizing those showing the strongest demographic differences. We select questions that exhibit significant variation across demographic groups to ensure the benchmark captures meaningful differences in responses. For each selected question, we create multiple test cases by pairing it with different values of the grouping variables (e.g., age = "18-29", age = "30-49"). This process results in 6,343 test cases that specifically measure LLMs' ability to simulate responses from narrower participant groups based on specified demographic characteristics. Table 3 provides a summary of the sampling process across all datasets, showing the minimum group size thresholds and the number of test cases in each benchmark split.

# D Metric Robustness Check

TVD ranges from 0 (perfect match) to 1 (complete disagreement), with lower values indicating better simulation fidelity. TVD provides an interpretable measure of how closely model predictions align with actual human response distributions. TVD is particularly well-suited for simulation evaluation compared to alternatives like KL divergence or Jensen-Shannon divergence (JSD). Unlike KL divergence, TVD remains well-defined even when the model assigns zero probability to responses that humans give, avoiding the infinite penalties that KL would impose in such cases. Additionally, TVD is symmetric and bounded, making it more interpretable across different datasets and response distributions than KL divergence. While JSD offers similar advantages in terms of symmetry and boundedness, TVD provides a more direct and intuitive interpretation of the maximum possible error in probability estimates. This property is especially valuable when evaluating how accurately models simulate the distribution of human responses rather than just matching the most likely response. For further discussion on TVD as an evaluation metric, see also [48]. We show the results of Table 1 in terms of raw TVD values in Table 5.

To ensure our findings are robust across different metrics, we complement TVD with two alternative metrics: Jensen-Shannon Divergence (JSD) and Spearman's Rank Correlation (RC). Table 4 presents these metrics for a subset of evaluated models. The strong Pearson correlation between TVD and JSD ($r = 0.92$) indicates these metrics provide consistent model rankings. The moderate negative correlation ($r = -0.57$) between TVD and RC is expected, as lower distances correspond to higher correlations. This multi-metric evaluation confirms that our model comparisons remain consistent across different statistical measures.

Table 3: Dataset Sampling Summary; NaN refers to dataset that is only available in aggregated form and no grouping size is known.

| Dataset | Min. Group | SimBench | SimBenchPop | SimBenchGrouped |
|---|---|---|---|---|
| WisdomOfCrowds | 100 | 1,604 | 114 | – |
| Jester | 100 | 136 | 136 | – |
| Choices13k | NaN | 14,568 | 500 | – |
| OpinionQA | 300 | 1,074,392 | 500 | 984 |
| MoralMachineClassic | 100 | 3,441 | 15 | – |
| MoralMachine | 100 | 20,771 | 500 | – |
| ChaosNLI | 100 | 4,645 | 500 | – |
| ESS | 300 | 2,783,780 | 500 | 1,643 |
| Afrobarometer | 300 | 517,453 | 500 | 1,531 |
| OSPsychBig5 | 300 | 1,950 | 250 | – |
| OSPsychMACH | 300 | 3,682,700 | 100 | – |
| OSPsychMGKT | 300 | 20,610 | 500 | – |
| OSPsychRWAS | 300 | 975,585 | 22 | – |
| ISSP | 300 | 594,336 | 500 | 940 |
| LatinoBarometro | 300 | 80,684 | 500 | 1,245 |
| GlobalOpinionQA | NaN | 46,329 | 500 | – |
| DICES | 10 | 918,064 | 500 | – |
| NumberGame | 10 | 15,984 | 500 | – |
| ConspiracyCorr | 300 | 968 | 45 | – |
| TISP | 300 | 172,271 | 485 | – |
| **Total** | | **10,930,271** | **7,167** | **6,343** |

Table 4: Comparison of models on three metrics: Total Variation Distance (TVD), Jensen-Shannon Divergence (JSD), and Spearman Rank Correlation (RC). Lower values are better for TVD and JSD; higher is better for RC.

| Model | Total Variation | JS Divergence | Rank Correlation |
|---|---|---|---|
| Claude-3.7-Sonnet | 0.191 | 0.057 | 0.673 |
| Claude-3.7-Sonnet-4000 | 0.195 | 0.060 | 0.648 |
| DeepSeek-R1 | 0.211 | 0.069 | 0.623 |
| DeepSeek-V3-0324 | 0.216 | 0.069 | 0.620 |
| GPT-4.1 | 0.209 | 0.070 | 0.646 |
| Llama-3.1-405B-Instruct | 0.231 | 0.085 | 0.593 |
| o4-mini-high | 0.225 | 0.079 | 0.621 |
| o4-mini-low | 0.230 | 0.082 | 0.609 |

# E Regression Analysis of Human Response Entropy and Model Performance

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:

$$\text{Total\_Variation} \sim C(\text{dataset\_name}) + C(\text{model}) + C(\text{instruct\_flag}) : \text{Human\_Normalized\_Entropy} \tag{2}$$

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 6 are the coefficients for the interaction terms:

- For base models: The coefficient on the interaction between base models and Human Normalized Entropy is $-0.2555$ ($p < 0.001$), indicating that for every one-unit increase in normalized entropy, the TVD decreases by approximately 0.26 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 $+0.1072$ ($p < 0.001$), 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 20% of the variance in TVD ($R^2 = 0.202$), 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.

Table 5: TVD for each model in SimBenchPop and SimBenchGrouped. Lower values indicate better performance. PT and IT refer to pretrained and instruction-tuned versions, respectively.

| **Model** | **SimBenchPop** | **SimBenchGrouped** | **Average** |
|---|---|---|---|
| *Baselines* | | | |
| Random baseline | 0.390 | 0.415 | 0.402 |
| Uniform baseline | 0.335 | 0.362 | 0.348 |
| *Commercial Models* | | | |
| Claude-3.7-Sonnet | 0.197 | 0.184 | 0.191 |
| Claude-3.7-Sonnet-4000 | 0.201 | 0.188 | 0.195 |
| GPT-4.1 | 0.212 | 0.205 | 0.209 |
| o4-mini-high | 0.235 | 0.214 | 0.225 |
| o4-mini-low | 0.234 | 0.216 | 0.230 |
| *Open Models* | | | |
| DeepSeek-V3-0324 | 0.215 | 0.218 | 0.216 |
| DeepSeek-R1 | 0.211 | 0.212 | 0.211 |
| Llama-3.1-8B-Instruct | 0.321 | 0.318 | 0.320 |
| Llama-3.1-70B-Instruct | 0.277 | 0.247 | 0.263 |
| Llama-3.1-405B-Instruct | 0.237 | 0.225 | 0.231 |
| Qwen2.5-0.5B | 0.337 | 0.364 | 0.349 |
| Qwen2.5-1.5B | 0.321 | 0.324 | 0.322 |
| Qwen2.5-3B | 0.300 | 0.327 | 0.313 |
| Qwen2.5-7B | 0.290 | 0.326 | 0.307 |
| Qwen2.5-14B | 0.285 | 0.314 | 0.298 |
| Qwen2.5-32B | 0.273 | 0.308 | 0.290 |
| Qwen2.5-72B | 0.269 | 0.300 | 0.283 |
| Gemma-3-1B-PT | 0.382 | 0.413 | 0.396 |
| Gemma-3-4B-PT | 0.334 | 0.342 | 0.338 |
| Gemma-3-12B-PT | 0.310 | 0.317 | 0.314 |
| Gemma-3-27B-PT | 0.309 | 0.325 | 0.317 |
| Gemma-3-4B-IT | 0.337 | 0.341 | 0.339 |
| Gemma-3-12B-IT | 0.262 | 0.274 | 0.267 |
| Gemma-3-27B-IT | 0.270 | 0.273 | 0.272 |

Table 6: Results: Ordinary least squares

| | | | |
|---|---|---|---|
| Model: | OLS | Adj. R-squared: | 0.201 |
| Dependent Variable: | Total_Variation | AIC: | -134342.8438 |
| Date: | 2025-05-15 20:27 | BIC: | -133890.3555 |
| No. Observations: | 172008 | Log-Likelihood: | 67216. |
| Df Model: | 44 | F-statistic: | 983.5 |
| Df Residuals: | 171963 | Prob (F-statistic): | 0.00 |
| R-squared: | 0.201 | Scale: | 0.026805 |

| | Coef. | Std.Err. | t | P> \|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 0.1824 | 0.0029 | 62.1882 | 0.0000 | 0.1766 | 0.1881 |
| C(dataset_name)[T.ChaosNLI] | -0.0442 | 0.0021 | -20.7195 | 0.0000 | -0.0483 | -0.0400 |
| C(dataset_name)[T.Choices13k] | -0.1016 | 0.0021 | -47.3233 | 0.0000 | -0.1058 | -0.0974 |
| C(dataset_name)[T.ConspiracyCorr] | -0.0452 | 0.0052 | -8.6565 | 0.0000 | -0.0554 | -0.0349 |
| C(dataset_name)[T.DICES] | -0.0254 | 0.0023 | -11.0298 | 0.0000 | -0.0300 | -0.0209 |
| C(dataset_name)[T.ESS] | -0.0202 | 0.0021 | -9.4882 | 0.0000 | -0.0244 | -0.0160 |
| C(dataset_name)[T.GlobalOpinionQA] | -0.0428 | 0.0021 | -20.2041 | 0.0000 | -0.0469 | -0.0386 |
| C(dataset_name)[T.ISSP] | -0.0279 | 0.0021 | -13.1516 | 0.0000 | -0.0321 | -0.0238 |
| C(dataset_name)[T.Jester] | 0.1168 | 0.0033 | 35.9190 | 0.0000 | 0.1104 | 0.1232 |
| C(dataset_name)[T.LatinoBarometro] | -0.0325 | 0.0021 | -15.1931 | 0.0000 | -0.0367 | -0.0283 |
| C(dataset_name)[T.MoralMachine] | -0.0380 | 0.0021 | -17.8607 | 0.0000 | -0.0422 | -0.0339 |
| C(dataset_name)[T.MoralMachineClassic] | -0.1594 | 0.0088 | -18.1961 | 0.0000 | -0.1766 | -0.1422 |
| C(dataset_name)[T.NumberGame] | -0.0821 | 0.0021 | -38.8471 | 0.0000 | -0.0863 | -0.0780 |
| C(dataset_name)[T.OSPsychBig5] | -0.1186 | 0.0026 | -45.0783 | 0.0000 | -0.1238 | -0.1134 |
| C(dataset_name)[T.OSPsychMACH] | -0.0227 | 0.0037 | -6.1522 | 0.0000 | -0.0299 | -0.0155 |
| C(dataset_name)[T.OSPsychMGKT] | -0.1121 | 0.0021 | -52.6066 | 0.0000 | -0.1163 | -0.1080 |
| C(dataset_name)[T.OSPsychRWAS] | 0.0168 | 0.0073 | 2.3068 | 0.0211 | 0.0025 | 0.0311 |
| C(dataset_name)[T.OpinionQA] | -0.1013 | 0.0021 | -47.9196 | 0.0000 | -0.1054 | -0.0972 |
| C(dataset_name)[T.TISP] | -0.0441 | 0.0022 | -20.5072 | 0.0000 | -0.0483 | -0.0399 |
| C(dataset_name)[T.WisdomOfCrowds] | -0.0200 | 0.0035 | -5.7228 | 0.0000 | -0.0268 | -0.0131 |
| C(Model)[T.Claude-3.7-Sonnet-4000] | 0.0038 | 0.0027 | 1.3978 | 0.1622 | -0.0015 | 0.0092 |
| C(Model)[T.DeepSeek-R1] | 0.0133 | 0.0027 | 4.8513 | 0.0000 | 0.0079 | 0.0186 |
| C(Model)[T.DeepSeek-V3-0324] | 0.0177 | 0.0027 | 6.4740 | 0.0000 | 0.0123 | 0.0231 |
| C(Model)[T.GPT-4.1] | 0.0141 | 0.0027 | 5.1557 | 0.0000 | 0.0087 | 0.0195 |
| C(Model)[T.Gemma-3-12B-IT] | 0.0641 | 0.0027 | 23.4327 | 0.0000 | 0.0587 | 0.0694 |
| C(Model)[T.Gemma-3-12B-PT] | 0.3616 | 0.0035 | 104.5204 | 0.0000 | 0.3549 | 0.3684 |
| C(Model)[T.Gemma-3-1B-PT] | 0.4330 | 0.0035 | 125.1390 | 0.0000 | 0.4262 | 0.4398 |
| C(Model)[T.Gemma-3-27B-IT] | 0.0730 | 0.0027 | 26.6890 | 0.0000 | 0.0676 | 0.0784 |
| C(Model)[T.Gemma-3-27B-PT] | 0.3604 | 0.0035 | 104.1666 | 0.0000 | 0.3536 | 0.3672 |
| C(Model)[T.Gemma-3-4B-IT] | 0.1398 | 0.0027 | 51.1034 | 0.0000 | 0.1344 | 0.1451 |
| C(Model)[T.Gemma-3-4B-PT] | 0.3857 | 0.0035 | 111.4826 | 0.0000 | 0.3790 | 0.3925 |
| C(Model)[T.Llama-3.1-405B-Instruct] | 0.0392 | 0.0027 | 14.3206 | 0.0000 | 0.0338 | 0.0445 |
| C(Model)[T.Llama-3.1-70B-Instruct] | 0.0792 | 0.0027 | 28.9426 | 0.0000 | 0.0738 | 0.0845 |
| C(Model)[T.Llama-3.1-8B-Instruct] | 0.1231 | 0.0027 | 45.0170 | 0.0000 | 0.1178 | 0.1285 |
| C(Model)[T.Qwen2.5-0.5B] | 0.3880 | 0.0035 | 112.1256 | 0.0000 | 0.3812 | 0.3947 |
| C(Model)[T.Qwen2.5-1.5B] | 0.3719 | 0.0035 | 107.4976 | 0.0000 | 0.3652 | 0.3787 |
| C(Model)[T.Qwen2.5-14B] | 0.3359 | 0.0035 | 97.0893 | 0.0000 | 0.3292 | 0.3427 |
| C(Model)[T.Qwen2.5-32B] | 0.3248 | 0.0035 | 93.8707 | 0.0000 | 0.3180 | 0.3316 |
| C(Model)[T.Qwen2.5-3B] | 0.3517 | 0.0035 | 101.6583 | 0.0000 | 0.3450 | 0.3585 |
| C(Model)[T.Qwen2.5-72B] | 0.3198 | 0.0035 | 92.4342 | 0.0000 | 0.3130 | 0.3266 |
| C(Model)[T.Qwen2.5-7B] | 0.3409 | 0.0035 | 98.5348 | 0.0000 | 0.3342 | 0.3477 |
| C(Model)[T.o4-mini-high] | 0.0374 | 0.0027 | 13.6575 | 0.0000 | 0.0320 | 0.0427 |
| C(Model)[T.o4-mini-low] | 0.0363 | 0.0027 | 13.2773 | 0.0000 | 0.0310 | 0.0417 |
| C(instruct_flag)[base]:Human_Normalized_Entropy | -0.2628 | 0.0026 | -101.0841 | 0.0000 | -0.2679 | -0.2577 |
| C(instruct_flag)[instruct]:Human_Normalized_Entropy | 0.0929 | 0.0024 | 37.9507 | 0.0000 | 0.0881 | 0.0977 |

| | | | |
|---|---|---|---|
| Omnibus: | 21133.651 | Durbin-Watson: | 1.711 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 34296.360 |
| Skew: | 0.862 | Prob(JB): | 0.000 |
| Kurtosis: | 4.346 | Condition No.: | 33 |

# F 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 (§2.2).

## F.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**: [61]

### F.2 Jester

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

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

**Example question**:

> How funny is the following joke, on a scale of -10 to 10? (-10: not funny, 10: very funny)
>
> How many feminists does it take to screw in a light bulb? That's not funny.
>
> Options:
> (A): 7 to 10
> (B): 3 to 6
> (C): -2 to 2
> (D): -5 to -3
> (E): -10 to -6

**Participants**: 7,669 volunteer participants (sociodemographics unknown) who chose to use the Jester joke recommender website.

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

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

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

**Publication**: [26]

### F.3 Choices13k

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

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

**Example question**:

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

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

**Participant grouping variables**: None.

**Default System Prompt**:

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

**License**: "All data are available to the public without registration at github.com/jcpeterson/choices13k".

**Publication**: [58]

### F.4 OpinionQA

**Description**:

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 repurposed to evaluate LLM alignment with the opinions of different sociodemographic groups.

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

**Example question**:

> How would you describe your household's financial situation?
>
> (A): Live comfortably
> (B): Meet your basic expenses with a little left over for extras
> (C): Just meet your basic expenses
> (D): Don't even have enough to meet basic expenses
> (E): Refused

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

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

**Default System Prompt**:

> You are from the United States.

**License**: No licensing information provided; Data is freely available without registration at `https://worksheets.codalab.org/worksheets/0x6fb693719477478aac73fc07db333f69`

**Publication**: [60]

### F.5 MoralMachineClassic

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

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

**Example question**:

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

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

**Participant grouping variables** (n=6): *country*: respondent's country of residence, *gender*: gender of the respondent, *education*: level of education, *age_group*: age bracket, *political_group*: self-identified political orientation, *religious_group*: self-identified religious affiliation.

**Default System Prompt**:

> The Moral Machine website (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.

**License**: No licensing information provided.

**Publication**: [9]

### F.6 ChaosNLI

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

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

**Example question**:

> Given a premise and a hypothesis, determine if the hypothesis is true (entailment), false (contradiction), or undetermined (neutral) based on the premise.
>
> Premise: Two young children in blue jerseys, one with the number 9 and one with the number 2 are standing on wooden steps in a bathroom and washing their hands in a sink.
> Hypothesis: Two kids at a ballgame wash their hands.
>
> Choose the most appropriate relationship between the premise and hypothesis:
> (A): Entailment (the hypothesis must be true if the premise is true)
> (B): Contradiction (the hypothesis cannot be true if the premise is true)
> (C): Neutral (the hypothesis may or may not be true given the premise)

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

**Participant grouping variables**: None.

**Default System Prompt**:

> You are an Amazon Mechanical Turk worker.

**License**: CC BY-NC 4.0

**Publication**: [50]

### F.7 European Social Survey (ESS)

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

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

**Example question**:

> Sometimes the government disagrees with what most people think is best for the country. Which one of the statements on this card describes what you think is best for democracy in general?
>
> Options:
> (A): Government should change its policies
> (B): Government should stick to its policies
> (C): It depends on the circumstances

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

**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.)

**Default System Prompt**:

> The year is {survey year}.

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

**Publication**: [25]

### F.8 AfroBarometer

**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.

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

**Example question**:

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

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

**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,

**Default System Prompt**:

The year is {survey year}.

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

**Publication**: [1]

### F.9 OSPsychBig5

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

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

**Example question**:

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

Options:
(A): Disagree
(B): Slightly Disagree
(C): Neutral
(D): Slightly Agree
(E): Agree

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

**Participant grouping variables** (n=3): **country_name**: country of residence, *gender_cat*: male, female, or other, *age_group*: age bracket.

**Default System Prompt**:

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

**License**: Creative Commons.

**Publication**: None.

### F.10 OSPsychMGKT

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

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

**Example question**:

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

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

**Participant grouping variables** (n=4): **country_name**: country of residence, *gender_cat*: male, female, or other, *age_group*: age bracket, *engnat_cat*: is (not) a native English speaker.

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

**License**: Creative Commons.

**Publication**: None.

### F.11 OSPsychMACH

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

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

**Example question**:

> Indicate your level of agreement with the following statement:
> Never tell anyone the real reason you did something unless it is useful to do so.
>
> Options:
> (A): Disagree
> (B): Slightly disagree
> (C): Neutral
> (D): Slightly agree
> (E): Agree

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

**Participant grouping variables** (n=18): **country_name**: country of residence, *gender_cat*: male, female, or other, *age_group*: age bracket, *race_cat*: respondent’s race, *engnat_cat*: is (not) a native English speaker, *hand_cat*: right-, left-, or both-handed, *education_cat*: level of education, *urban_cat*: type of urban area, *religion_cat*: stated religion, *orientation_cat*: sexual orientation, *voted_cat*: did (not) vote at last elections, *married_cat*: never, currently, or previously married, *familysize*: number of people belonging to the family, *TIPI_E_Group*: extraversion level based on TIPI score, *TIPI_A_Group*: agreeableness level based on TIPI score, *TIPI_C_Group*: conscientiousness level based on TIPI score, *TIPI_ES_Group*: emotional stability level based on TIPI score, *TIPI_O_Group*: openness-to-experience level based on TIPI score.

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

**License**: Creative Commons.

**Publication**: None.

### F.12 OSPsychRWAS

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

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

**Example question**:

> Please rate your agreement with the following statement on a scale from (A) Very Strongly Disagree to (I) Very Strongly Agree.
>
> Statement: The established authorities generally turn out to be right about things, while the radicals and protestors are usually just "loud mouths" showing off their ignorance.
>
> 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

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

**Participant grouping variables** (n=18): *age_group*: age bracket, *gender_cat*: male or female or other, *race_cat*: respondent's race, *engnat_cat*: is (not) English native, *hand_cat*: right/left/both-handed, *education_cat*: level of education, *urban_cat*: type of urban area, *religion_cat*: stated religion, *orientation_cat*: sexual orientation, *voted*: did (not) vote at last elections, *married*: never/currently/previously, *familysize*: number of people belonging to the family, *TIPI_E_Group*: extraversion level based on TIPI score, *TIPI_A_Group*: agreeableness level based on TIPI score, *TIPI_C_Group*: conscientiousness level based on TIPI score, *TIPI_ES_Group*: emotional stability level based on TIPI score, *TIPI_O_Group*: openness-to-experience level based on TIPI score. *household_income*: income sufficiency, *work_status*: job situation, *religion*: stated religion, *nr_of_persons_in_household*: 1-7+, *marital_status* respondent's legal relationship status, *domicil*: type of urban area,

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

**License**: Creative Commons.

**Publication**: None.

### F.13 International Social Survey Programme (ISSP)

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

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

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

**Participant grouping variables** (n=11): *country*: respondent's country, *age*: age bracket, *gender*: male or female, *years_of_education*: 1-10+, *household_income*: income sufficiency, *work_status*: job situation, *religion*: stated religion, *nr_of_persons_in_household*: 1-7+, *marital_status* respondent's legal relationship status, *domicil*: type of urban area, *topbot*: self-asessed social class

**Default System Prompt**:

> The timeframe is {survey timeframe}.

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

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

#### F.13.1 ISSP 2017 Social Networks and Social Resources

**Example question**:

> This section is about who you would turn to for help in different situations, if you needed it.
>
> For each of the following situations, please tick one box to say who you would turn to first. If there are several people you are equally likely to turn to, please tick the box for the one you feel closest to.
>
> Who would you turn to first to help you around your home if you were sick and had to stay in bed for a few days?
>
> Options:
> (A): Close family member
> (B): More distant family member
> (C): Close friend
> (D): Neighbour
> (E): Someone I work with
> (F): Someone else
> (G): No one
> (H): Can't choose

**Publication**: [34]

#### F.13.2 ISSP 2018 Religion IV

**Example question**:

> Please indicate which statement below comes closest to expressing what you believe about God.
>
> Options:
> (A): I don't believe in God
> (B): Don't know whether there is a God and no way to find out
> (C): Don't believe in a personal God, but in a Higher Power
> (D): Find myself believing in God sometimes, but not at others
> (E): While I have doubts, I feel that I do believe in God
> (F): I know God really exists and have no doubts about it
> (G): Don't know

**Publication**: [35]

#### F.13.3 ISSP 2019 Social Inequality V

**Example question**:

> Looking at the list below, who do you think should have the greatest responsibility for reducing differences in income between people with high incomes and people with low incomes?
>
> Options:
> (A): Cant choose
> (B): Private companies
> (C): Government
> (D): Trade unions
> (E): High-income individuals themselves
> (F): Low-income individuals themselves
> (G): Income differences do not need to be reduced

**Publication**: [36]

#### F.13.4 ISSP 2020 Environment IV

**Example question**:

> In the last five years, have you ...
>
> Taken part in a protest or demonstration about an environmental issue?
>
> Options:
> (A): Yes, I have
> (B): No, I have not

**Publication**: [37]

#### F.13.5 ISSP 2021 Health and Health Care II

**Example question**:

> During the past 12 months, how often, if at all, have you used the internet to look for information on the following topics?
>
> Information related to anxiety, stress, or similar problems?
>
> Options:
> (A): Can't choose
> (B): Never
> (C): Seldom
> (D): Sometimes
> (E): Often
> (F): Very often

**Publication**: [38]

### F.14 LatinoBarómetro

**Description**:

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

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

**Example question**:

> Generally speaking, would you say you are satisfied with your life? Would you say you are...
>
> (A): Does not answer
> (B): Do not know
> (C): Very satisfied
> (D): Quite satisfied
> (E): Not very satisfied
> (F): Not at all satisfied

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

**Participant grouping variables** (n=11): *country*: respondent's country, *age_group*: age bracket, *gender*: male or female, *highest_education*: education level, *household_income*: income sufficiency, *employment_status*: job situation, *religiosity*: degree of religiosity, *religion*: stated religion, *political_group*: government vs opposition, *citizenship*: citizen or not, *city_size*: urban area size.

**Default System Prompt**:

> The year is {survey year}.

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

**Publication**: [43]

### F.15 GlobalOpinionQA

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

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

**Example question**:

> Question: Do you strongly agree, agree, disagree, or strongly disagree with the following statement:
> "On the whole, men make better business executives than women do."
>
> (A) Agree strongly
> (B) Agree
> (C) Disagree
> (D) Strongly disagree
> (E) Don't know

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

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

**Default System Prompt**: None.

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

**Publication**: [21]

### F.16 Diversity in Conversational AI Evaluation for Safety (DICES) - 990

**Description**: This dataset comprises 990 multi-turn adversarial conversations between humans and a dialogue model, accompanied by raters' **judgement** on the safety of last model response. This

dataset captures variability in human perception, making it especially valuable for analyzing how demographic diversity influences safety assessments of conversational AI.

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

**Example question**:

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

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

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

> You are a crowd worker.

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

**Publication**: [7]

### F.17 NumberGame

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

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

**Example question**:

> A program produces the following numbers: 63_ 43.
>
> Is it likely that the program generates this number next: 24?
> (A): Yes
> (B): No

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

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

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

**License**: CC0 1.0.

**Publication**: [10]

### F.18 ConspiracyCorr

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

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

**Example question**:

> Would you say the following statement is true or false?
>
> Statement: The US Government knowingly helped to make the 9/11 terrorist attacks happen in America on 11 September, 2001
>
> Options:
> (A): Definitely true
> (B): Probably true
> (C): Probably false
> (D): Definitely false
> (E): Don't know

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

**Participant grouping variables** (n=4): *Country*: country of origin, *Age_Group*: age bracket of the respondent, *Gender*: gender of the respondent, *Gender*: highest level of education completed

> The year is {survey year}.

**License**: CC0 1.0 Universal.

**Publication**: [22]

### F.19 MoralMachine

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

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

**Example question**:

> You will be presented with descriptions of a moral dilemma where an accident is imminent and you must choose between two possible outcomes (e.g., 'Stay Course' or 'Swerve'). Each outcome will result in different consequences. Which outcome do you choose?
>
> Options:
>
> (A): Stay, outcome: in this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in the death of the pedestrians.
> Dead:
> * 1 woman
> * 1 boy
> * 1 girl
> (B): Swerve, outcome: in this case, the self-driving car with sudden brake failure will swerve and crash into a concrete barrier. This will result in the death of the passengers.
> Dead:
> * 1 woman

> * 1 elderly man
> * 1 elderly woman

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

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

> The Moral Machine website (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.

**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.

**Publication**: [8]

### F.20 Trust in Science and Science-Related Populism (TISP)

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

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

**Example question**:

> How concerned or not concerned are most scientists about people's wellbeing?
>
> Options:
> (A): not concerned
> (B): somewhat not concerned
> (C): neither nor
> (D): somewhat concerned
> (E): very concerned

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

**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.

> The year is {survey year}.

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

**Publication**: [47]

## G Additional Related Work

**Distribution Elicitation Methodologies** Prior research has primarily relied on first token probabilities to obtain survey answers from LLMs [60, 19, 64]. 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 [63, 48]. Nevertheless, calibration of LLM outputs remains an open challenge; while extensively studied for model answer confidence [70, 39, 41, 71] and hallucinations [40], 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 [16].

**Simulation of Complex Human Behavior** Few recent works have investigated LLM capabilities for simulation of temporal changes in human behavior [44]. [3] 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. [56] 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.

# H Implementation Details

For base models, we use HuggingFace Transformers [67] to run inference on a single NVIDIA RTX A6000 Ada GPU. We structure prompts so that the next token corresponds to the model's answer choices. For models smaller than 70B parameters, we use 8-bit quantization implemented in bitsandbytes [18], while 70B models use 4-bit quantization.

For instruction-tuned models, we use API calls. OpenAI models are accessed directly through their API, while other models are accessed via OpenRouter. We request verbalized probability outputs in JSON format with temperature initially set to 0. If parsing fails, we increase temperature to 1 and retry up to 5 times. All models successfully produced valid JSON under these conditions. When probability outputs do not sum to 1, we apply normalization.

Our evaluation includes a diverse set of models: Qwen 2.5 [68] (0.5B-72B), Gemma 3 [62] PT and IT (4B-27B), o4-mini [54], Claude 3.7 Sonnet [5], DeepSeek R1 [28], DeepSeek-V3-0324 [17], GPT-4.1 [53], and Llama-3.1-Instruct (8B-405B) [49].

To ensure the validity of our results, we perform two checks: 1) We verify that base models assign the vast majority of probability mass to the provided answer options. Even for small models like Qwen2.5-0.5B, the sum of probabilities across answer tokens is as high as 0.98, confirming that models rarely predict tokens outside the designated answer space. 2) We also evaluate the effect of quantization on model performance using a subset of SimBench. As shown in Table 7, performance remains consistent across quantization levels, with minimal variation in total variation scores even for quantization-sensitive models like Llama-3.1.

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

The following system prompt was consistent across all experimental conditions:

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

For token probability prediction, we adapted the prompt structure from [51]:

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

Prompt for eliciting verbalized probability prediction:

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

Table 7: Total Variation for different models at various quantization levels. Lower values indicate 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 |

# NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction accurately reflect the contributions made and the results presented in Section 4.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations of this work are discussed in Appendix A.

   Guidelines:

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

limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: SimBench is permissively licensed and available on GitHub and Hugging Face. All information needed for reproduction are described in Section 2.

   Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: SimBench is permissively licensed and available on GitHub and Hugging Face. All information needed for reproduction are described in Section 2.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: All data splits are described in Section 2.3. All evaluated models are described in Section 3. Further implementation details are provided in Appendix H.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We report error bars where appropriate.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The compute ressources needed for the experiments are described in Appendix H.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Our work conforms with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

Justification: We discuss broader impacts and risk associated with simulating human behavior in Appendix B.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [NA]

Justification: Our paper does not pose such risk.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: We have included proper credit and license information of existing datasets we used in Section F.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: We have described the details of the creation process of SimBench in Section 2.

    Guidelines:

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

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

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

    Answer: [NA]

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

    Guidelines:

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

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

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

    Answer: [NA]

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

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [NA]

    Justification: We did not use LLM as an important, original or non-standard component of the core methods in this research.

    Guidelines:

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