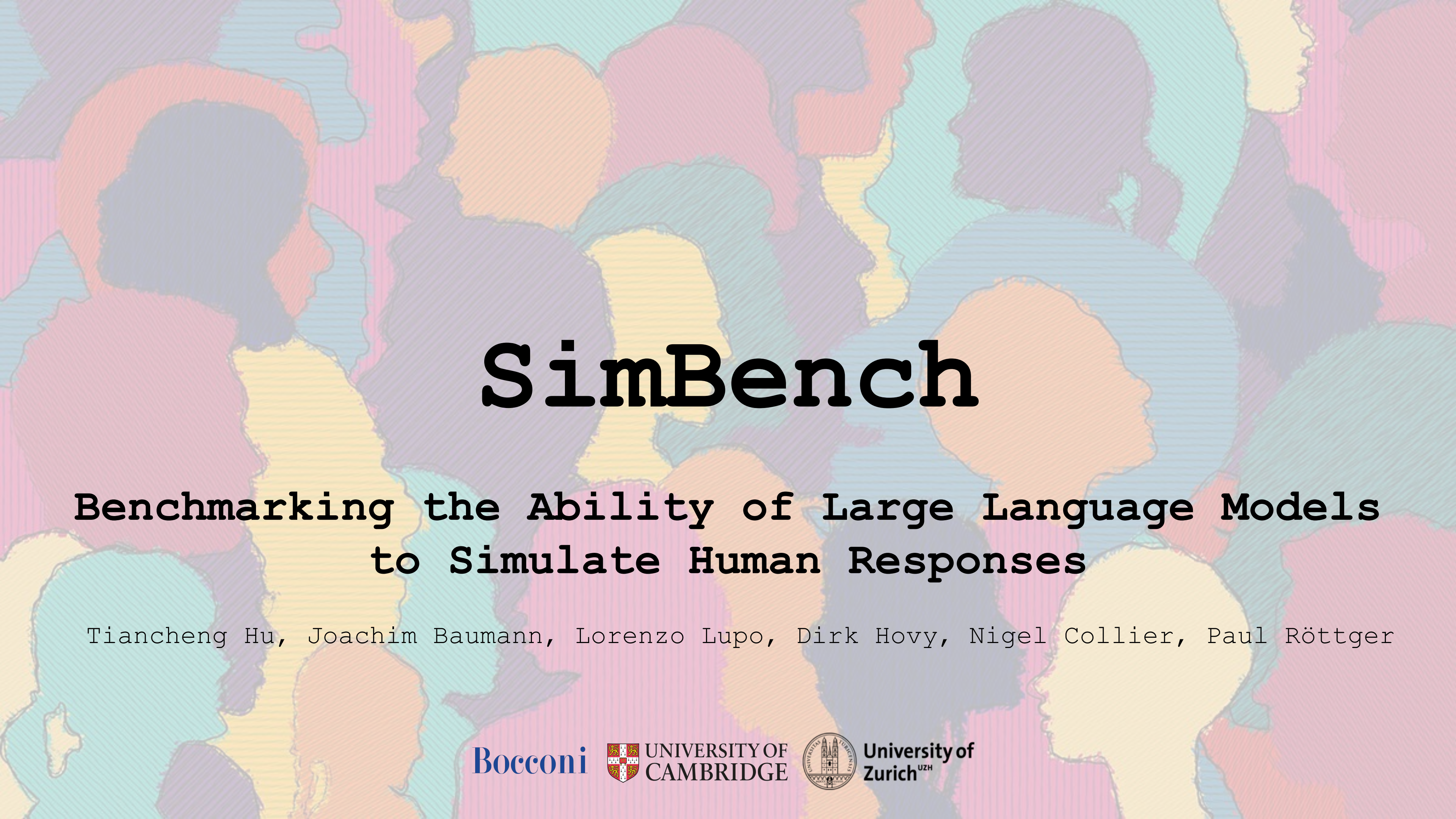


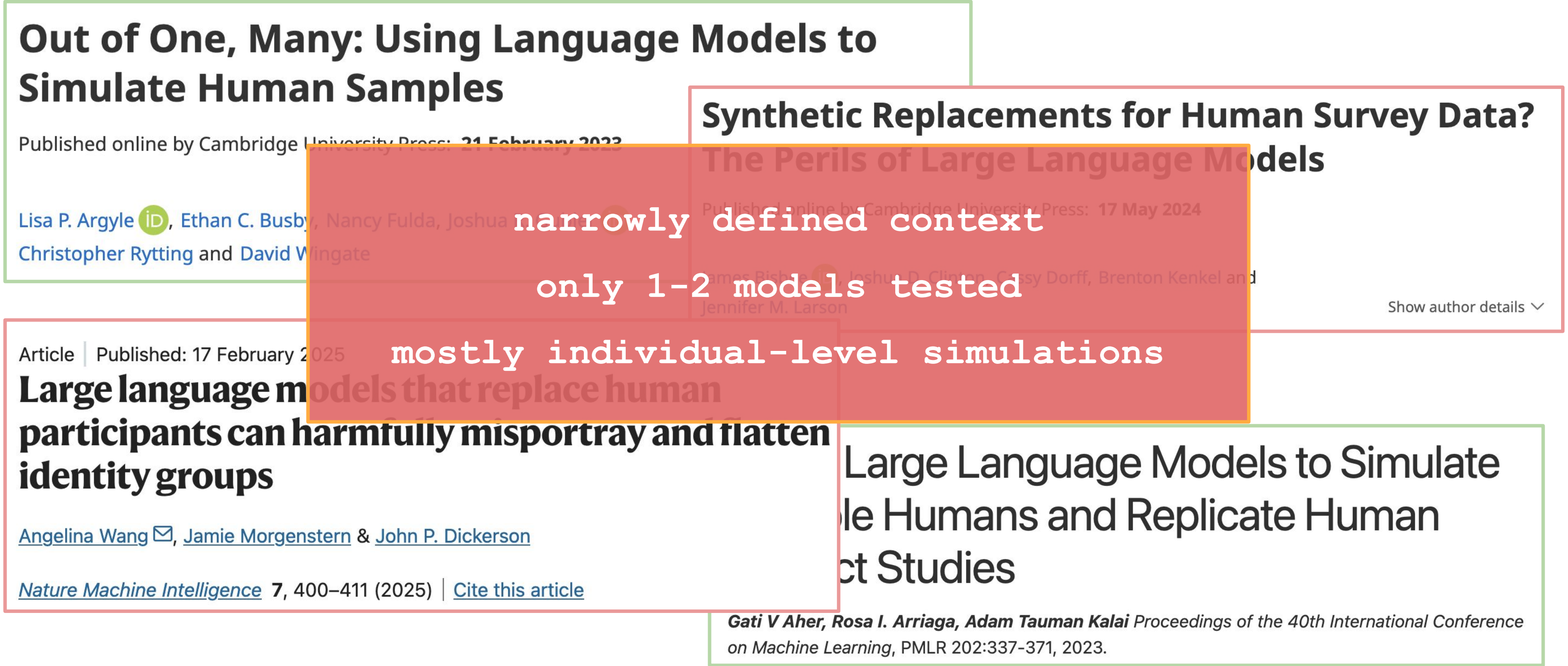
SimBench

Benchmarking the Ability of Large Language Models to Simulate Human Responses

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Dirk Hovy, Nigel Collier, Paul Röttger

Bocconi UNIVERSITY OF CAMBRIDGE University of Zurich^UZH

Prior Work



Out of One, Many: Using Language Models to Simulate Human Samples

Published online by Cambridge University Press: 21 February 2023

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua
Christopher Rytting and David Wingate

Synthetic Replacements for Human Survey Data? The Perils of Large Language Models

Published online by Cambridge University Press: 17 May 2024

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel and Jennifer M. Larson

Show author details

narrowly defined context

only 1-2 models tested

mostly individual-level simulations

Article | Published: 17 February 2025

Large language models that replace human participants can harmfully misportray and flatten identity groups

Angelina Wang, Jamie Morgenstern & John P. Dickerson

Nature Machine Intelligence **7**, 400–411 (2025) | Cite this article

Large Language Models to Simulate le Humans and Replicate Human ct Studies

Gati V Aher, Rosa I. Arriaga, Adam Tauman Kalai *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202:337-371, 2023.

Why simulating human responses?

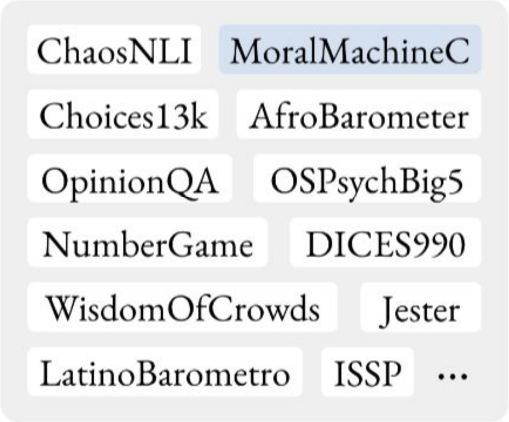
- **Replace** costly **surveys** and experiments
- **Pretest interventions** or public policies
- Explore **counterfactuals**

Why benchmarking LLMs?

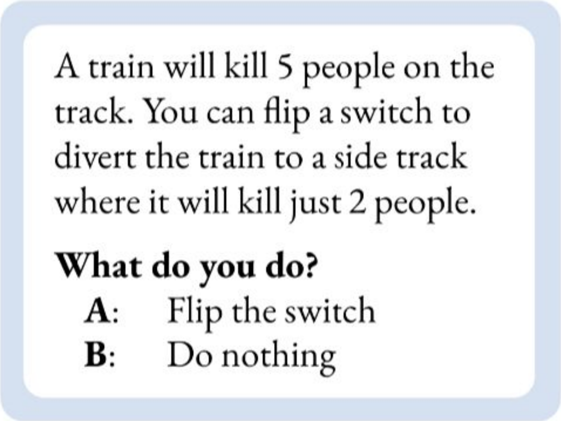
- How accurate simulations are?
- On which tasks?
- For which demographics?
- AI alignment

SimBench

20 datasets



ChaosNLI MoralMachineC
Choices13k AfroBarometer
OpinionQA OSPsychBig5
NumberGame DICES990
WisdomOfCrowds Jester
LatinoBarometro ISSP ···



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

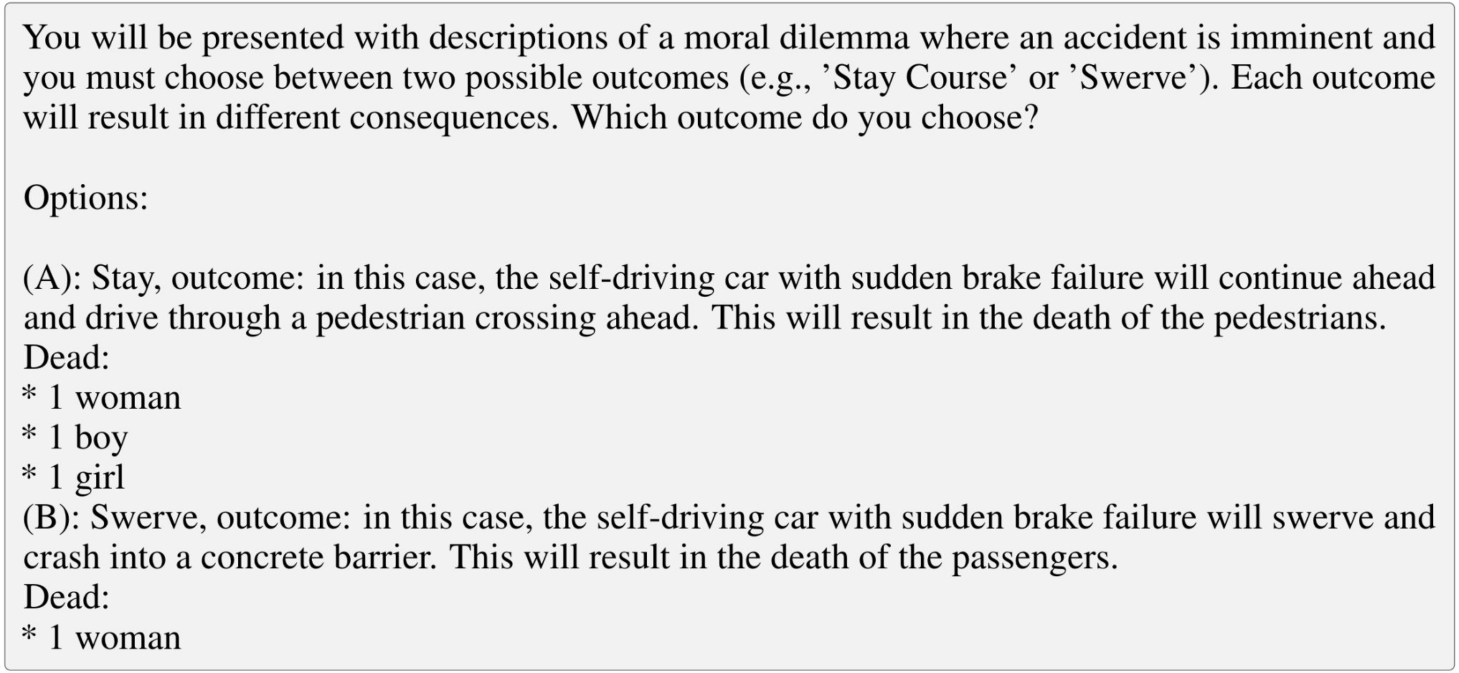
What do you do?

A: Flip the switch
B: Do nothing

Diverse tasks:

1. decision making
2. self-assessment
3. judgment
4. problem-solving

SimBench - *decision making*



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?

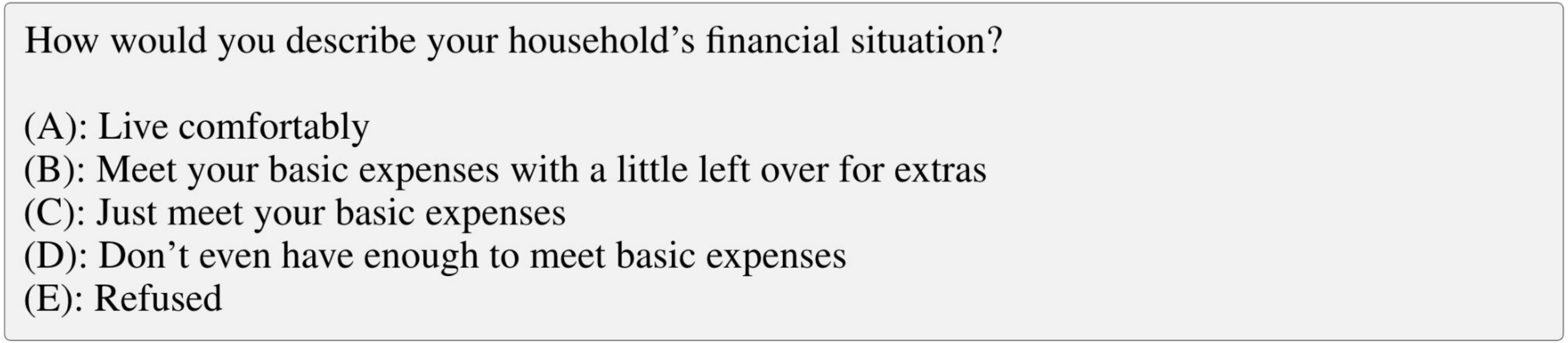
Options:

(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

MoralMachine

SimBench - *self-assessment*

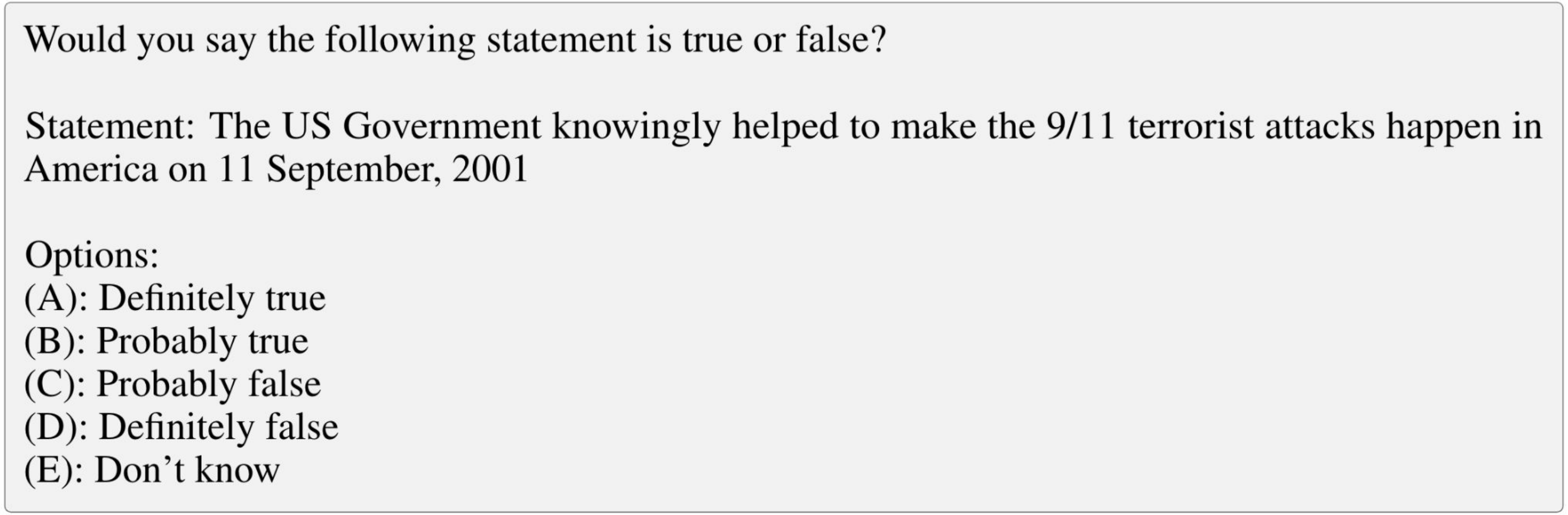


How would you describe your household’s financial situation?

(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

OpinionQA

SimBench - *judgment*



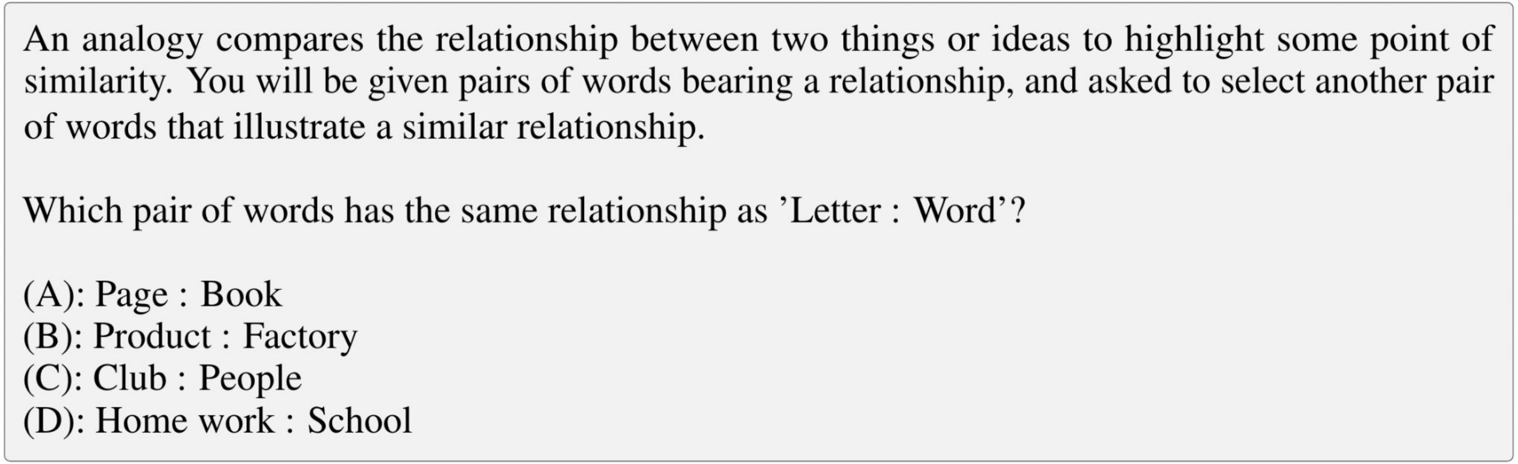
Would you say the following statement is true or false?

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

ConspiracyCorr

SimBench - *problem solving*



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.

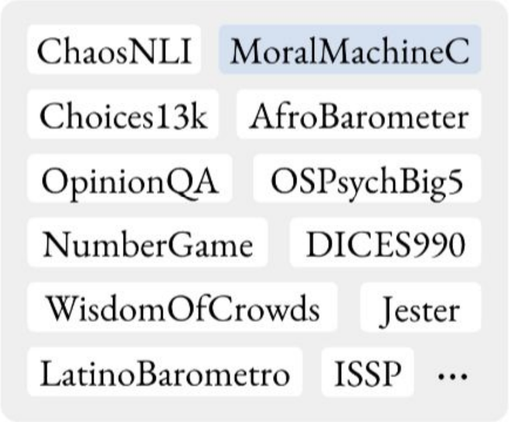
Which pair of words has the same relationship as ’Letter : Word’?

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

WisdomOfCrowds

SimBench

20 datasets



ChaosNLI MoralMachineC

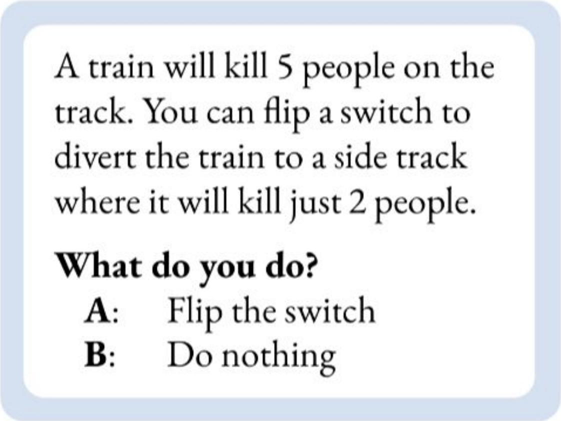
Choices13k AfroBarometer

OpinionQA OSPsychBig5

NumberGame DICES990

WisdomOfCrowds Jester

LatinoBarometro ISSP …



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

What do you do?

A: Flip the switch

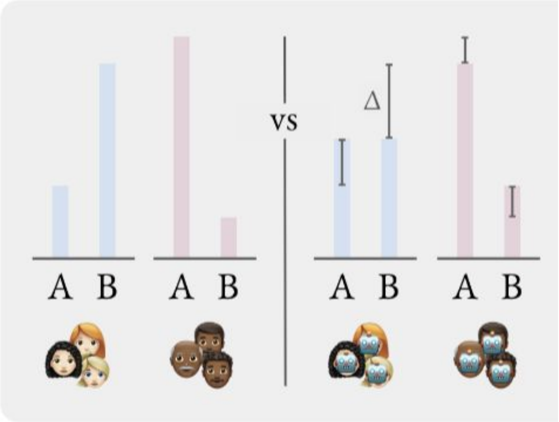
B: Do nothing

Diverse tasks:

1. decision making
2. self-assessment
3. judgment
4. problem-solving

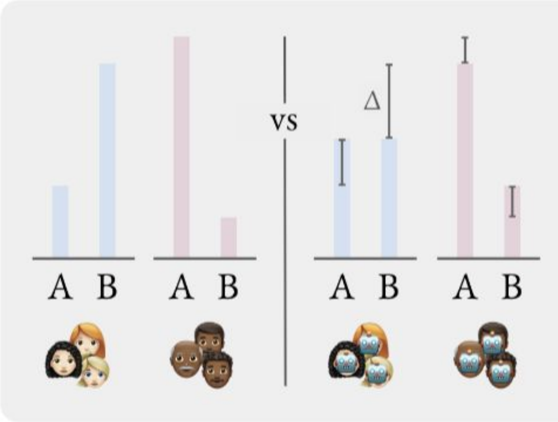
Participants diversity:

1. 130 countries
2. 6 continents
3. 8/20 representative

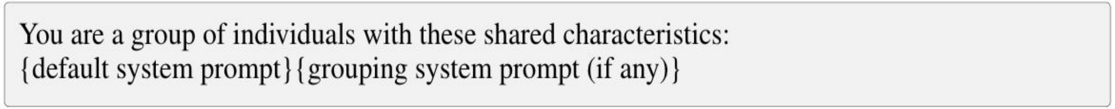


Experiments - Benchmarking LLMs on SimBench

- **RQ1:** How well do LLMs simulate human behavior across tasks?
- **RQ2:** How do model characteristics affect LLM simulation?
- **RQ3:** Do LLMs simulate all tasks equally well?
- **RQ4:** Do LLMs simulate all demographics equally well?

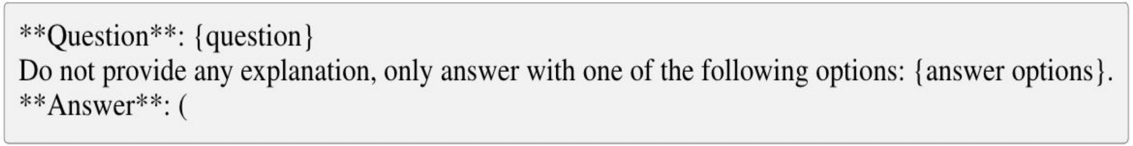


Experiments - LLM elicitation



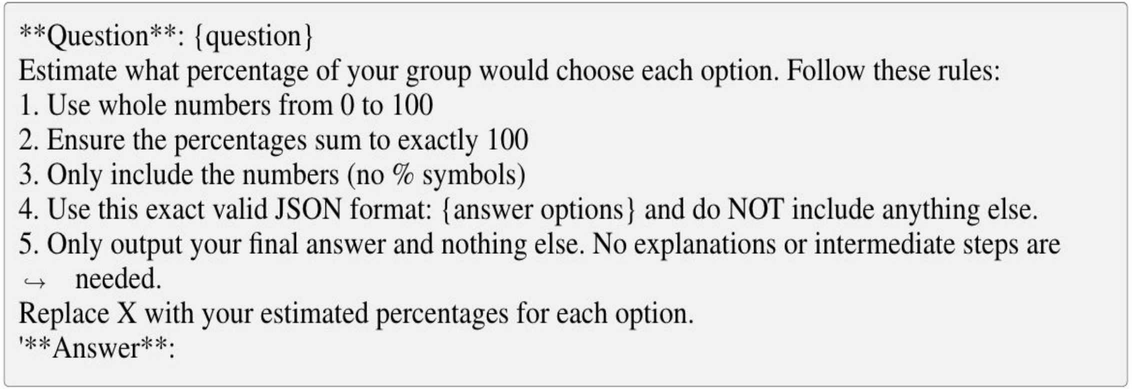
```
You are a group of individuals with these shared characteristics:
{default system prompt}{grouping system prompt (if any)}
```

METHOD 1: token probabilities

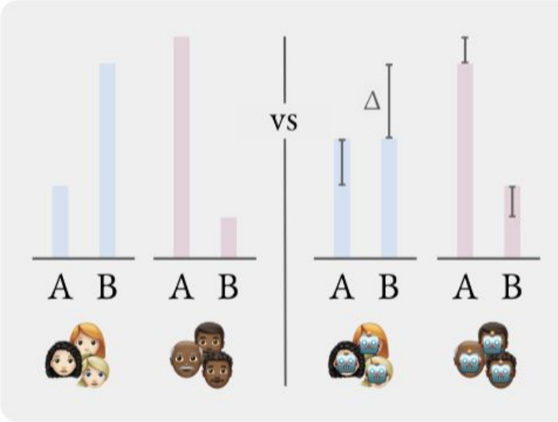


```
**Question**: {question}
Do not provide any explanation, only answer with one of the following options: {answer options}.
**Answer**: (
```

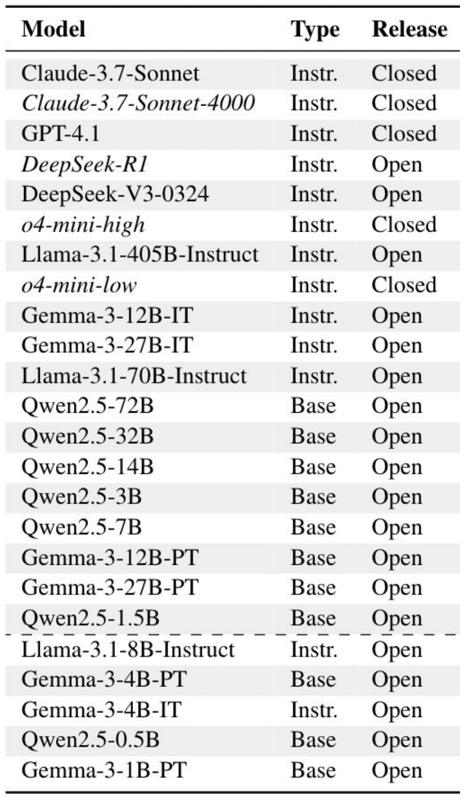
METHOD 2: verbalized probabilities



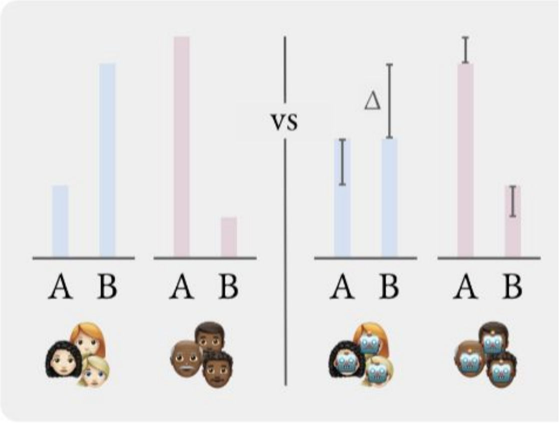
```
**Question**: {question}
Estimate what percentage of your group would choose each option. Follow these rules:
1. Use whole numbers from 0 to 100
2. Ensure the percentages sum to exactly 100
3. Only include the numbers (no % symbols)
4. Use this exact valid JSON format: {answer options} and do NOT 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**:
```



Experiments - Models

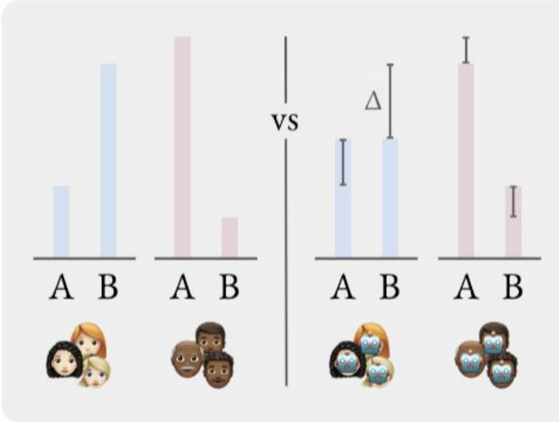


Model	Type	Release
Claude-3.7-Sonnet	Instr.	Closed
Claude-3.7-Sonnet-4000	Instr.	Closed
GPT-4.1	Instr.	Closed
DeepSeek-R1	Instr.	Open
DeepSeek-V3-0324	Instr.	Open
o4-mini-high	Instr.	Closed
Llama-3.1-405B-Instruct	Instr.	Open
o4-mini-low	Instr.	Closed
Gemma-3-12B-IT	Instr.	Open
Gemma-3-27B-IT	Instr.	Open
Llama-3.1-70B-Instruct	Instr.	Open
Qwen2.5-72B	Base	Open
Qwen2.5-32B	Base	Open
Qwen2.5-14B	Base	Open
Qwen2.5-3B	Base	Open
Qwen2.5-7B	Base	Open
Gemma-3-12B-PT	Base	Open
Gemma-3-27B-PT	Base	Open
Qwen2.5-1.5B	Base	Open
Llama-3.1-8B-Instruct	Instr.	Open
Gemma-3-4B-PT	Base	Open
Gemma-3-4B-IT	Instr.	Open
Qwen2.5-0.5B	Base	Open
Gemma-3-1B-PT	Base	Open

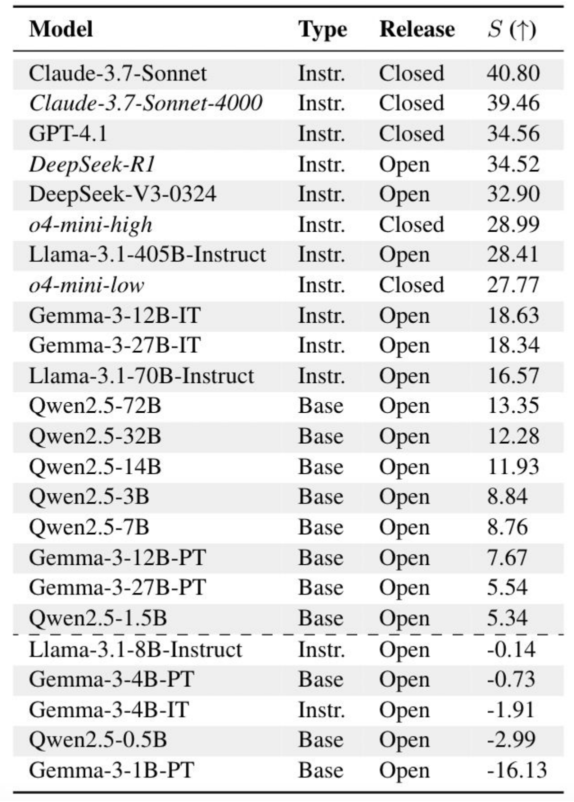


Experiments - Evaluation

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

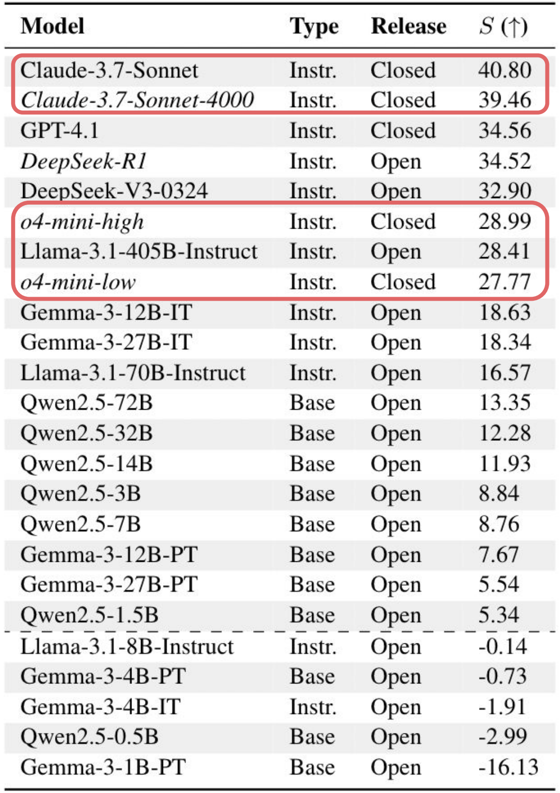


Results - RQ1 How well do LLMs simulate human behavior across tasks?

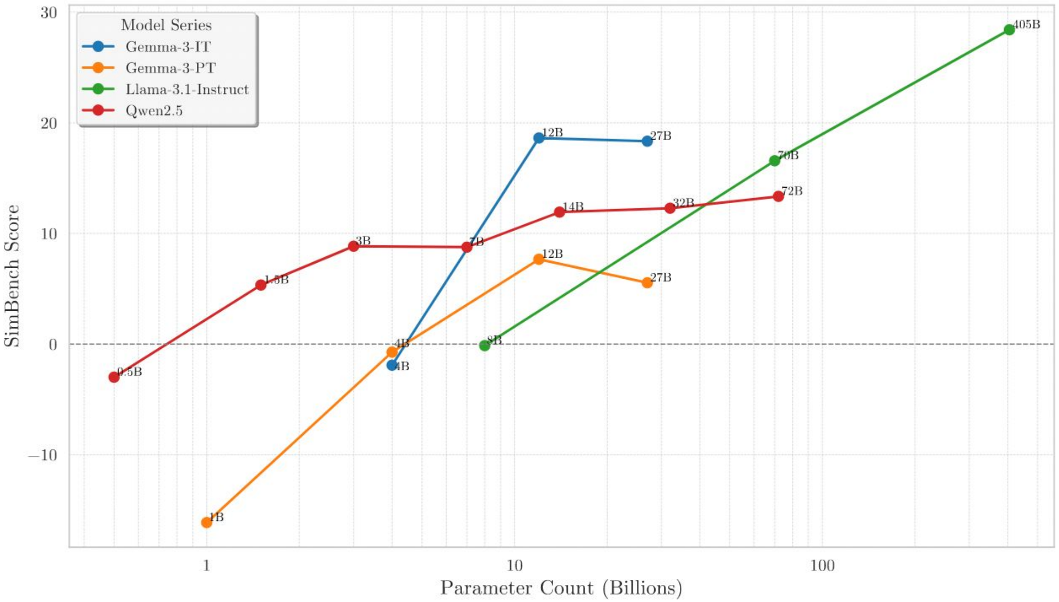


Model	**Type**	**Release**	S (↑)
Claude-3.7-Sonnet	Instr.	Closed	40.80
Claude-3.7-Sonnet-4000	Instr.	Closed	39.46
GPT-4.1	Instr.	Closed	34.56
DeepSeek-R1	Instr.	Open	34.52
DeepSeek-V3-0324	Instr.	Open	32.90
o4-mini-high	Instr.	Closed	28.99
Llama-3.1-405B-Instruct	Instr.	Open	28.41
o4-mini-low	Instr.	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

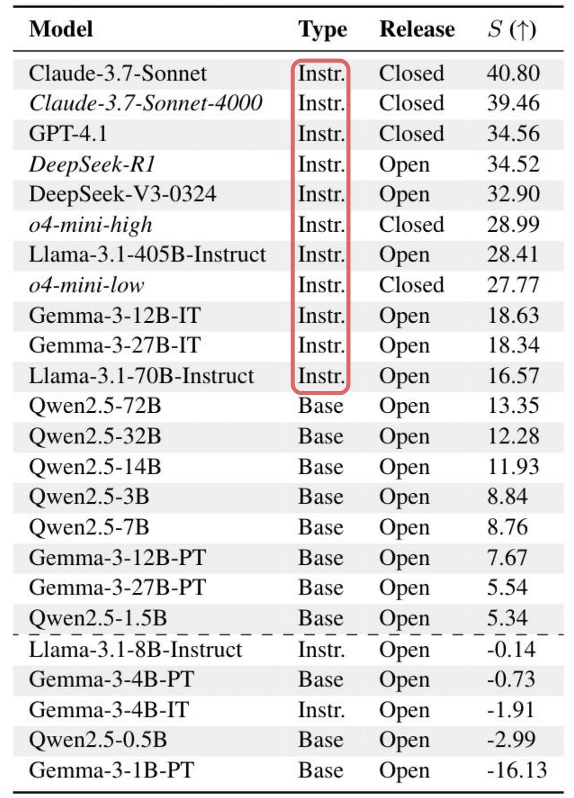
Results - RQ2 How do model characteristics affect LLM simulation?



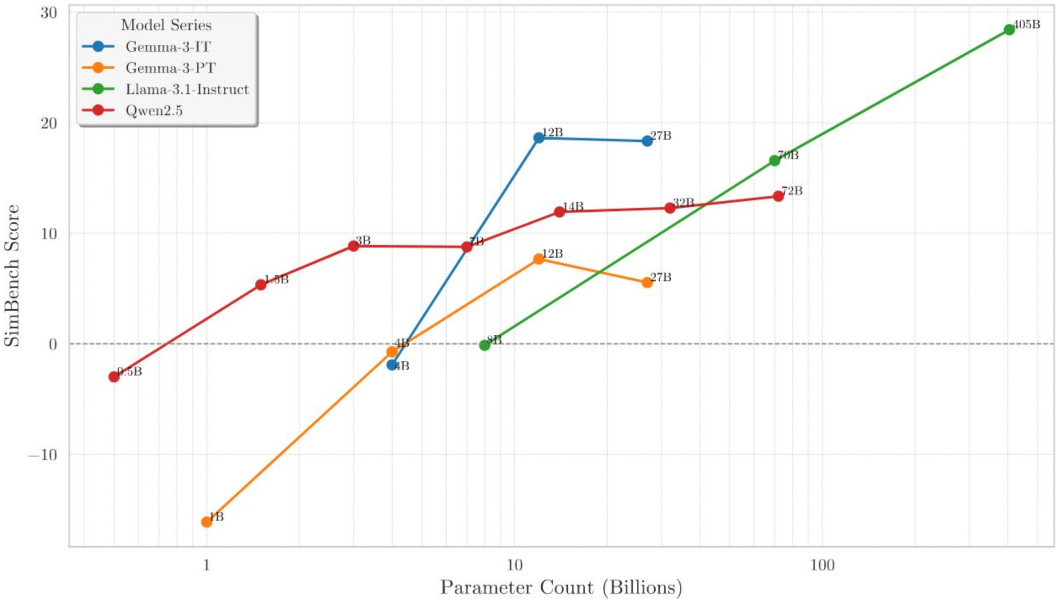
Model	Type	Release	S (↑)
Claude-3.7-Sonnet	Instr.	Closed	40.80
Claude-3.7-Sonnet-4000	Instr.	Closed	39.46
GPT-4.1	Instr.	Closed	34.56
DeepSeek-R1	Instr.	Open	34.52
DeepSeek-V3-0324	Instr.	Open	32.90
o4-mini-high	Instr.	Closed	28.99
Llama-3.1-405B-Instruct	Instr.	Open	28.41
o4-mini-low	Instr.	Closed	27.77
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



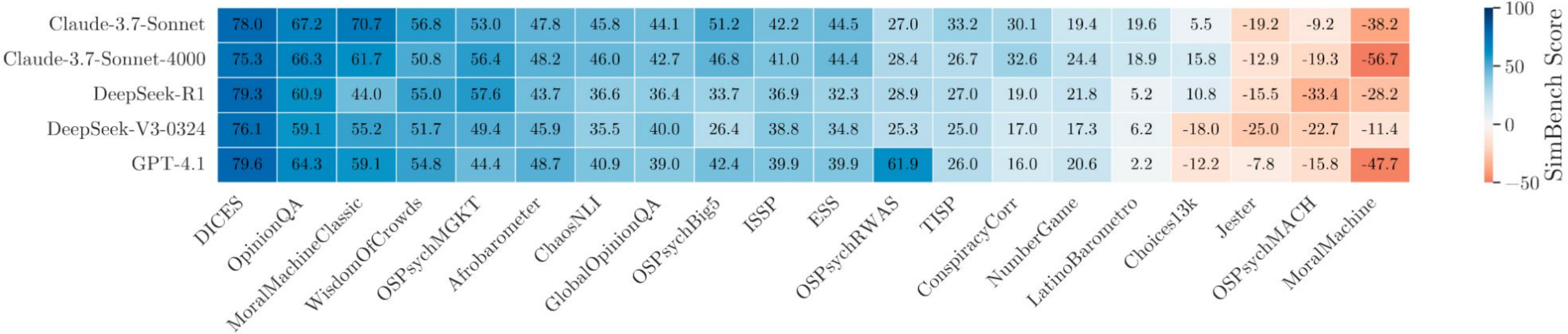
Results - RQ2 How do model characteristics affect LLM simulation?



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



Results - RQ3 Do LLMs simulate all tasks equally well?



	DICES	OpinionQA	MoralMachineClassic	WisdomOfCrowds	OSPsychMGKT	Afrobarometer	ChaosNLI	GlobalOpinionQA	OSPsychBig5	ISSP	ESS	OSPsychRWAS	TISP	ConspiracyCorr	NumberGame	LatinoBarometro	Choices13k	Jester	OSPsychMACH	MoralMachine
Claude-3.7-Sonnet	78.0	67.2	70.7	56.8	53.0	47.8	45.8	44.1	51.2	42.2	44.5	27.0	33.2	30.1	19.4	19.6	5.5	-19.2	-9.2	-38.2
Claude-3.7-Sonnet-4000	75.3	66.3	61.7	50.8	56.4	48.2	46.0	42.7	46.8	41.0	44.4	28.4	26.7	32.6	24.4	18.9	15.8	-12.9	-19.3	-56.7
DeepSeek-R1	79.3	60.9	44.0	55.0	57.6	43.7	36.6	36.4	33.7	36.9	32.3	28.9	27.0	19.0	21.8	5.2	10.8	-15.5	-33.4	-28.2
DeepSeek-V3-0324	76.1	59.1	55.2	51.7	49.4	45.9	35.5	40.0	26.4	38.8	34.8	25.3	25.0	17.0	17.3	6.2	-18.0	-25.0	-22.7	-11.4
GPT-4.1	79.6	64.3	59.1	54.8	44.4	48.7	40.9	39.0	42.4	39.9	39.9	61.9	26.0	16.0	20.6	2.2	-12.2	-7.8	-15.8	-47.7

Results - RQ4 Do LLMs simulate all demographics equally well?

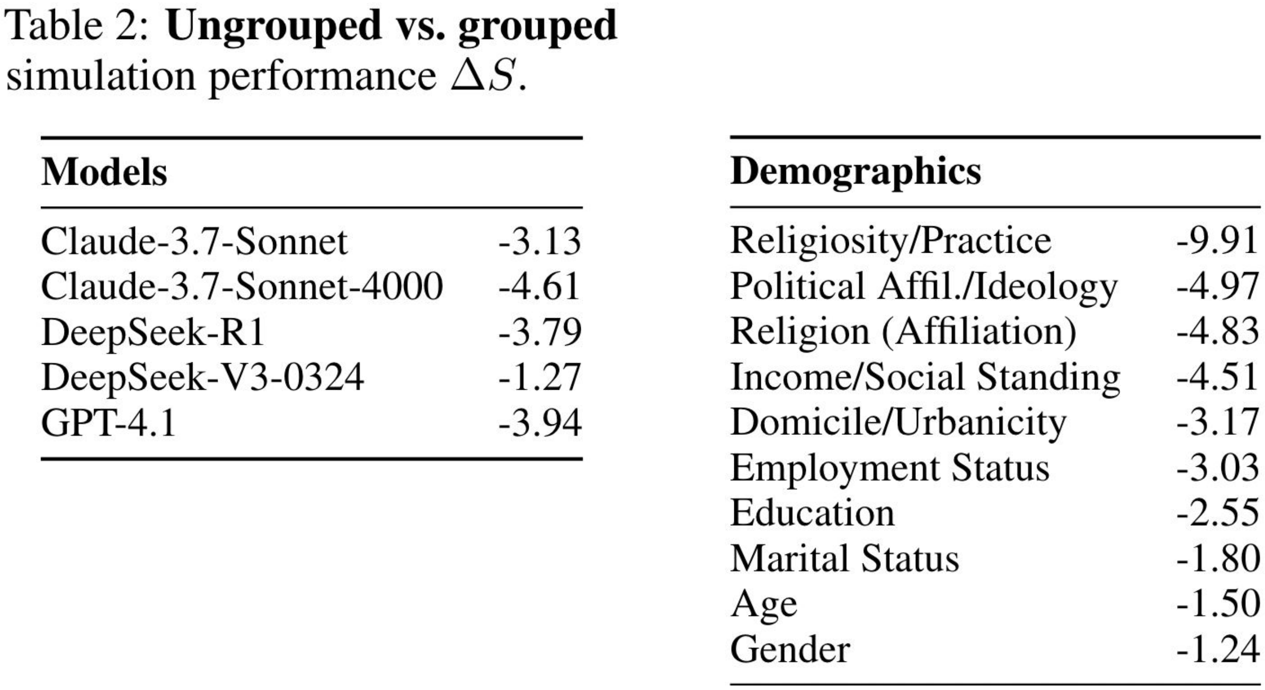


Table 2: **Ungrouped vs. grouped** simulation performance ΔS.

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

Wrapping Up

- Simulation with LLMs literature presents **mixed results** with:
  - only 1-2 models tested narrowly defined context
  - mostly individual-level simulations

- **Simbench - a benchmark for LLMs simulation capabilities**
  - large variety of tasks and respondents
  - group-level predictions

- **Experiments with 24 LLMs** showing that:
  - RQ1: LLMs are not great simulators
  - RQ2: Scaling laws make us hope for better simulators
  - RQ3: Disparate performance across tasks
  - RQ4: Disparate performance across demographic groups

Thanks for your attention!

lorelupo.github.io