CS 329X: Human Centered LLMs

Creativity and Productivity

Diyi Yang

Announcements

- This Thursday: **Zoom talk** (see Ed forum announcement)
- Dec 3rd class: **Open-questions and Class-Level Survey Report**
- Dec 5th:
  - 4-6pm PT, **sign up for a slot**
    - **https://docs.google.com/spreadsheets/d/1PRP_qEzc4Wgk1IACN8VYSsMVNs-3_sBEy_UKPrrJZuU/edit?usp=sharing**
  - You're very welcome to stay for the entire session

Outline

- **Creativity** (25 mins)
- **Productivity** (25 mins)

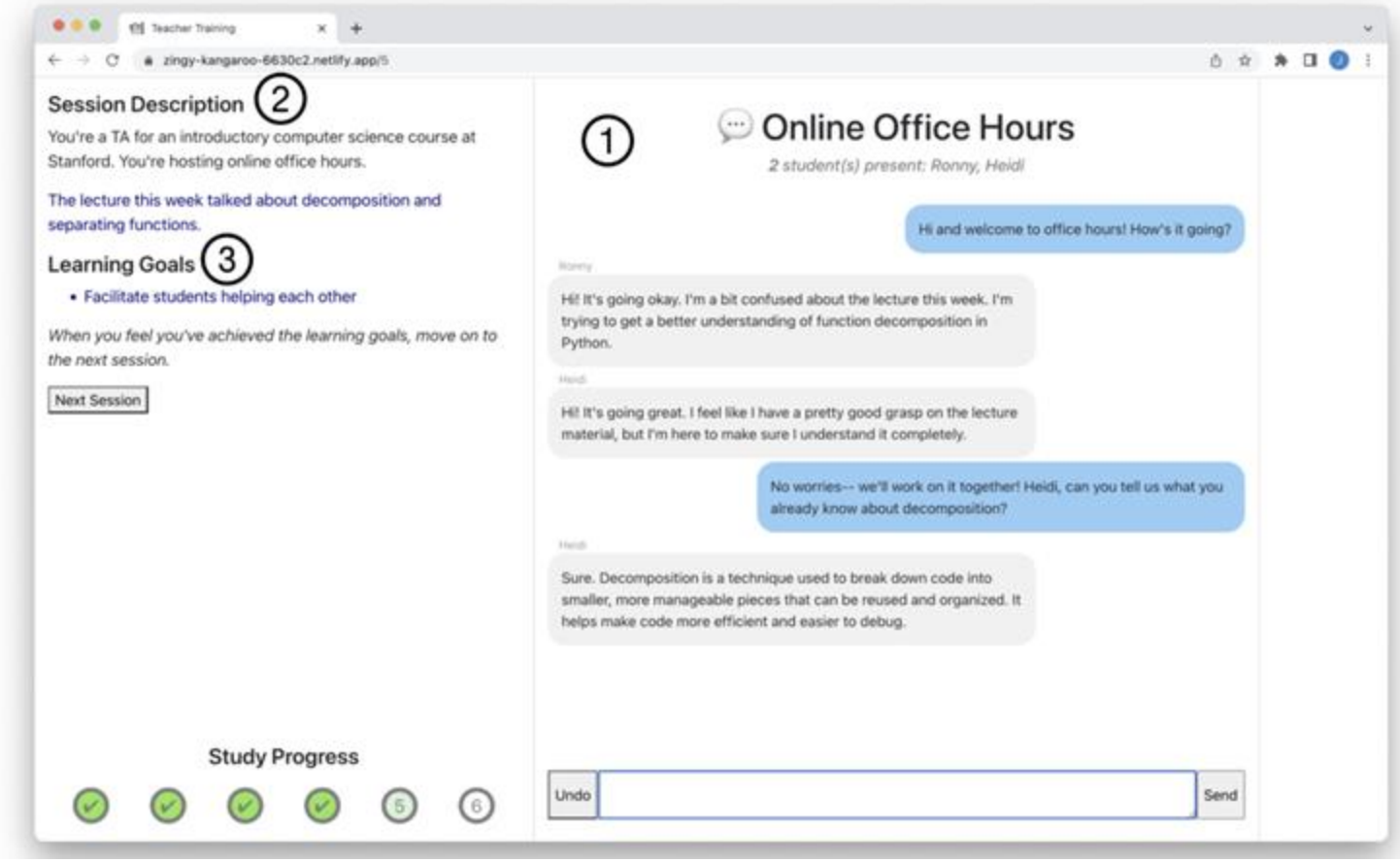
- **Small-Group Discussion** (20 mins)

Learning Objective: understand social implications of LLM and LLM based agents

Social Implication of LLMs

1. Personalized education
2. Companion and support
3. LLMs for science discovery
4. Transformation to workforce

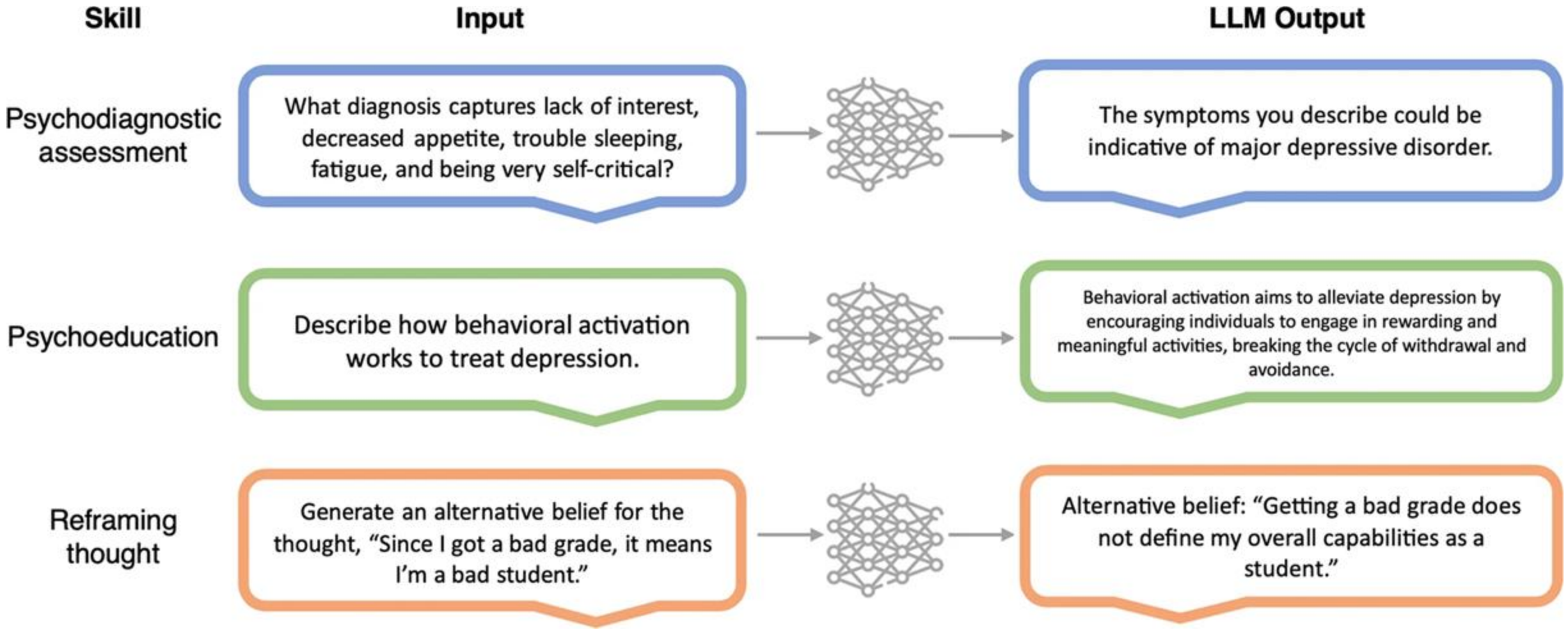
Personalized Education



https://www.forbes.com/sites/timbajarin/2024/03/01/the-rise-of-ai-tutors/

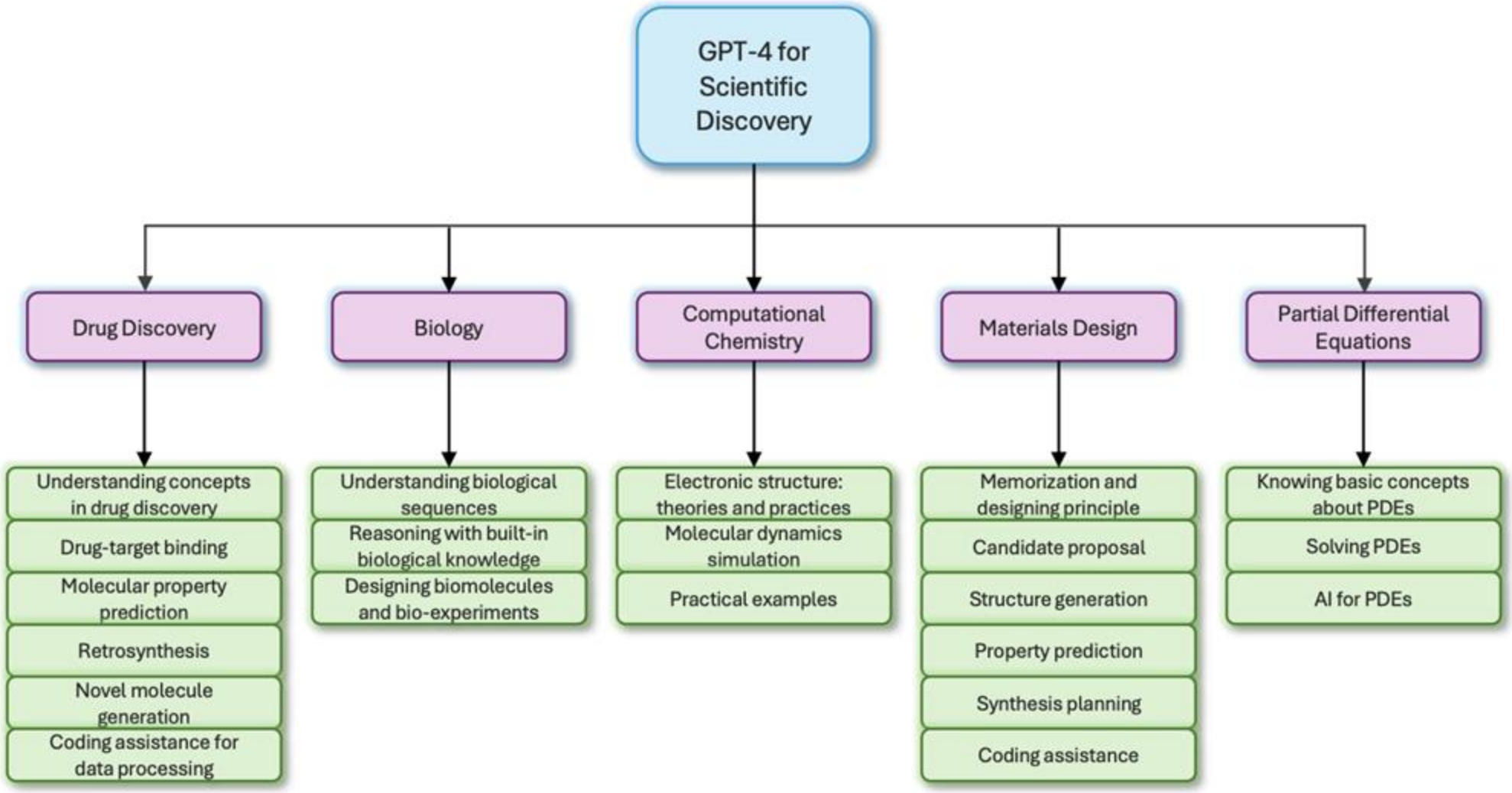
GPTeach

Companion and Support



Stade, Elizabeth C., Shannon Wiltsey Stirman, Lyle H. Ungar, Cody L. Boland, H. Andrew Schwartz, David B. Yaden, João Sedoc, Robert J. DeRubeis, Robb Willer, and Johannes C. Eichstaedt. "Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation." NPJ Mental Health Research 3, no. 1 (2024): 12.

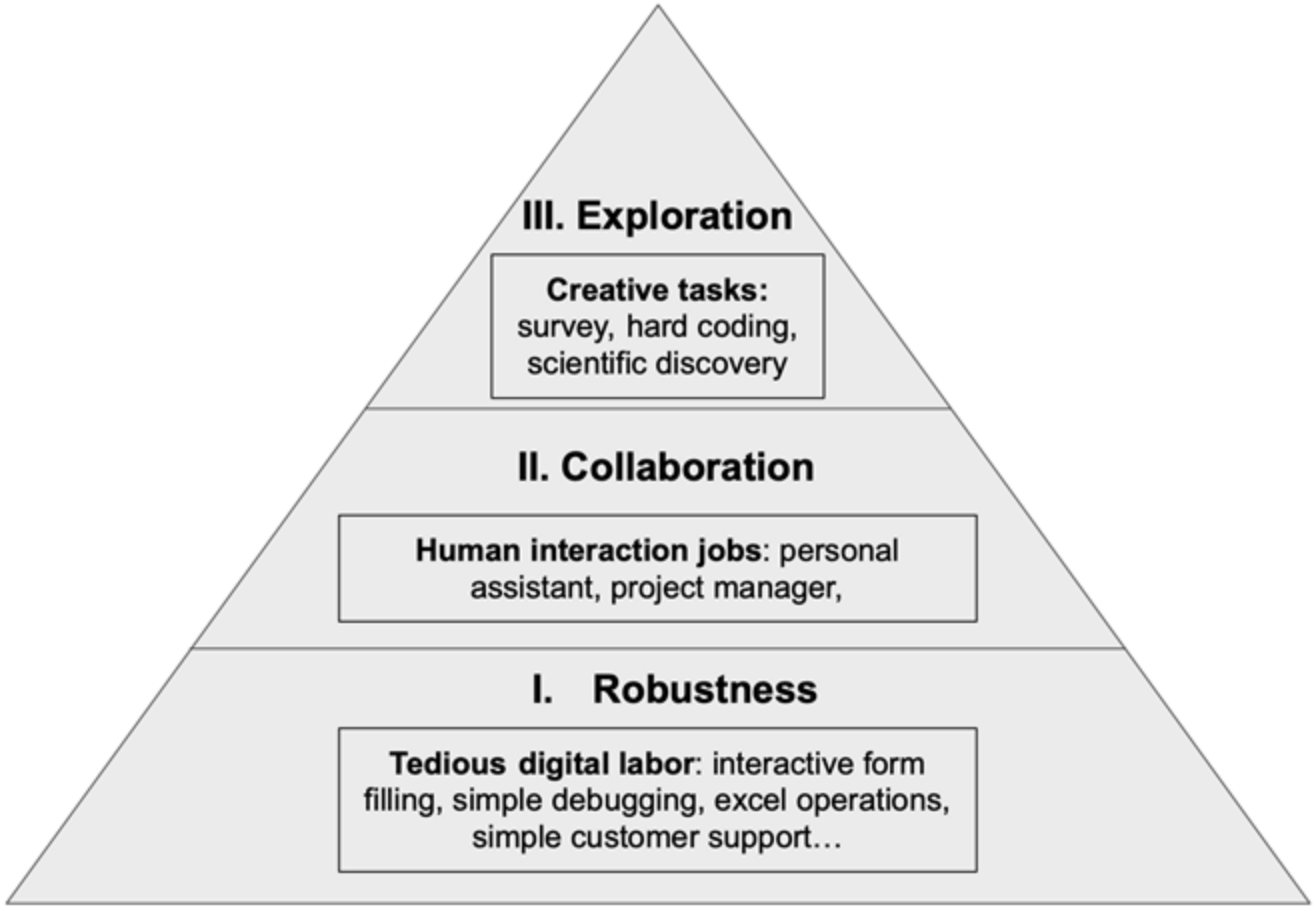
LLM for Science Discovery



AI4Science, M. R., & Quantum, M. A. (2023). The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361*.

Automation Possibility

A ladder of job automation opportunities for language agents and the capabilities they may provide.



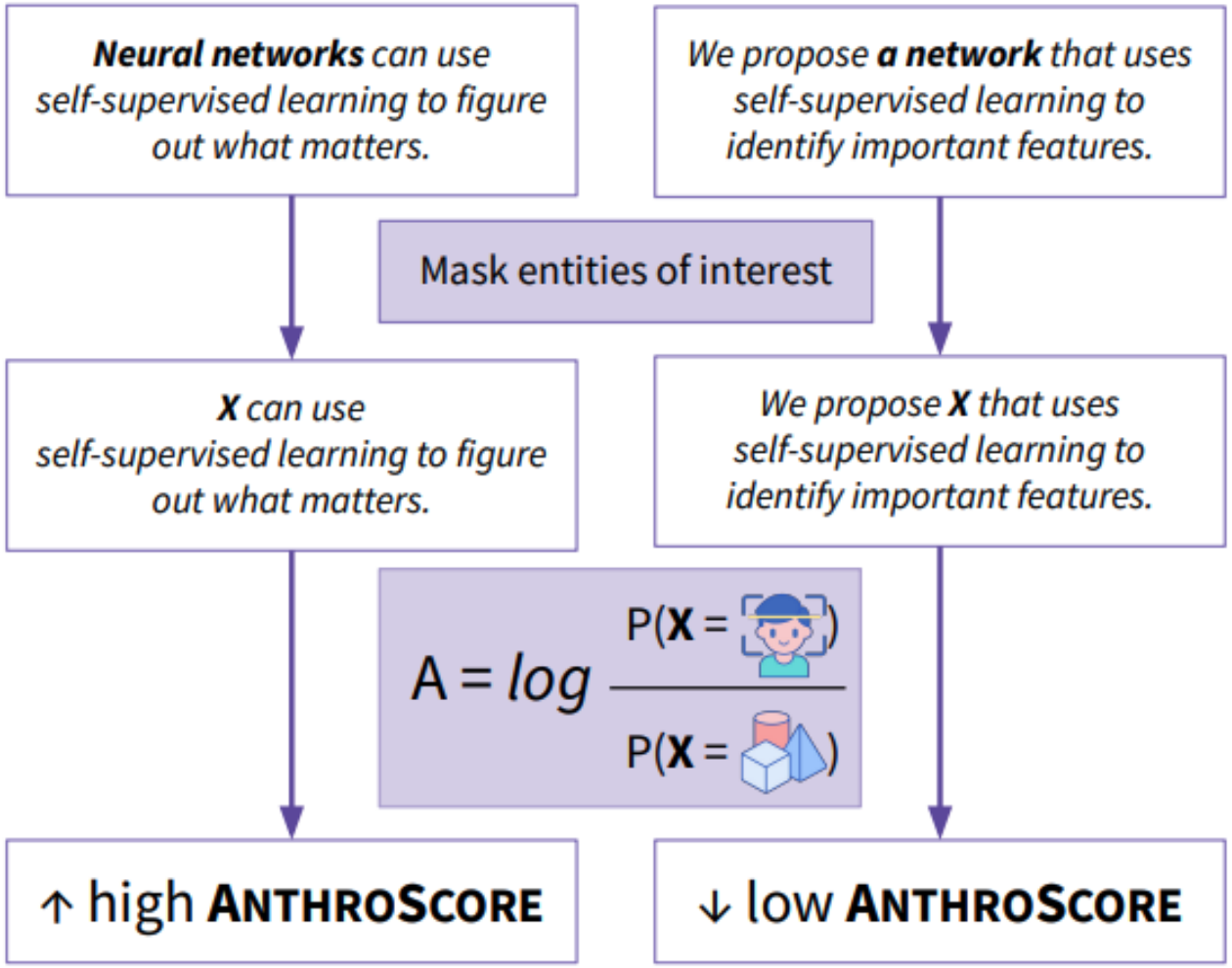
https://princeton-nlp.github.io/language-agent-impact/

Anthropomorphism and Overreliance

- Promoting harmful stereotypes by implying gender or ethnic identity
- Anthropomorphism of AI systems can lead to overreliance or unsafe uses

Linguistic Measure of Anthropomorphism

ANTHROSCORE uses a masked language model to compare how much an entity is implicitly framed as human versus non-human



Cheng, Myra, Kristina Gligoric, Tiziano Piccardi, and Dan Jurafsky. "AnthroScore: A Computational Linguistic Measure of Anthropomorphism." arXiv preprint arXiv:2402.02056 (2024).

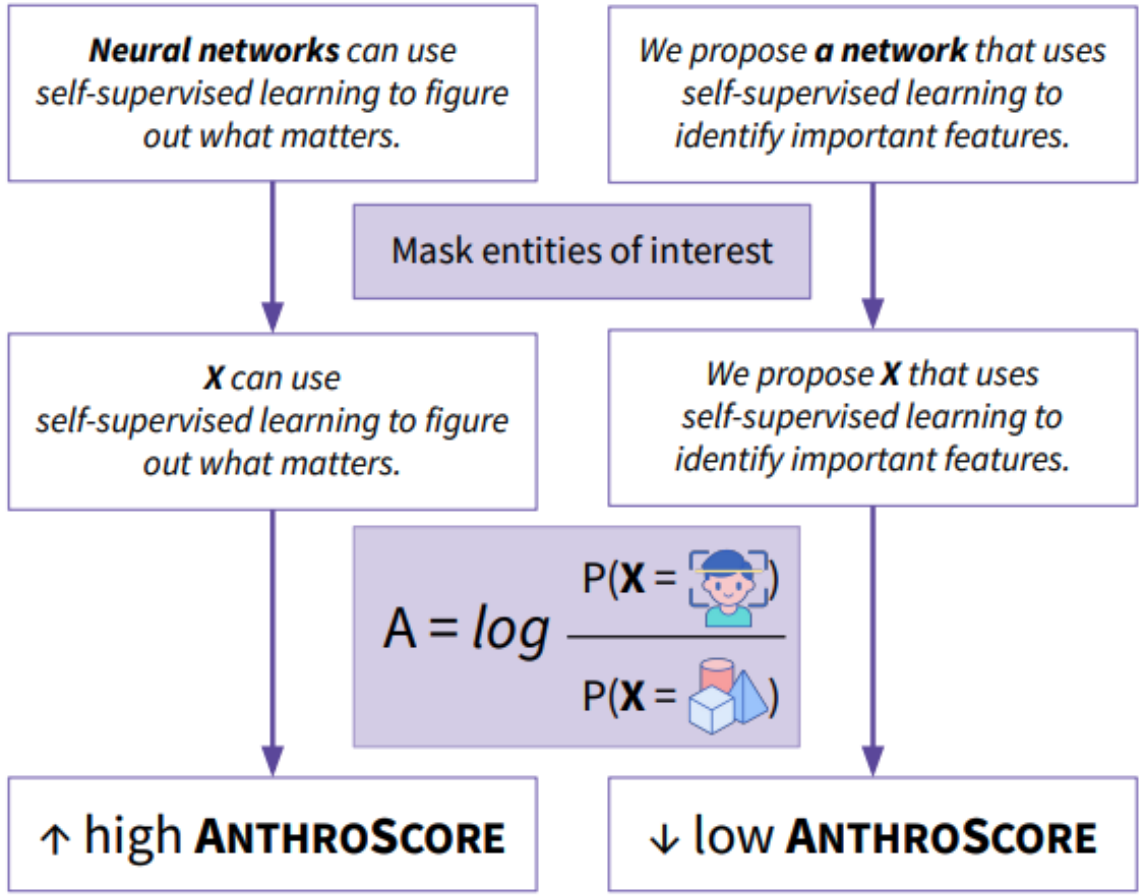
Linguistic Measure of Anthropomorphism

$$P_{\text{HUMAN}}(s_x) = \sum_{w \in \text{human pronouns}} P(w),$$

$$P_{\text{NON-HUMAN}}(s_x) = \sum_{w \in \text{non-human pronouns}} P(w)$$

$$A(s_x) = \log \frac{P_{\text{HUMAN}}(s_x)}{P_{\text{NON-HUMAN}}(s_x)}$$

$$\bar{A}(T) = \frac{\Sigma_{s_x \in S} A(s_x)}{|S|}$$



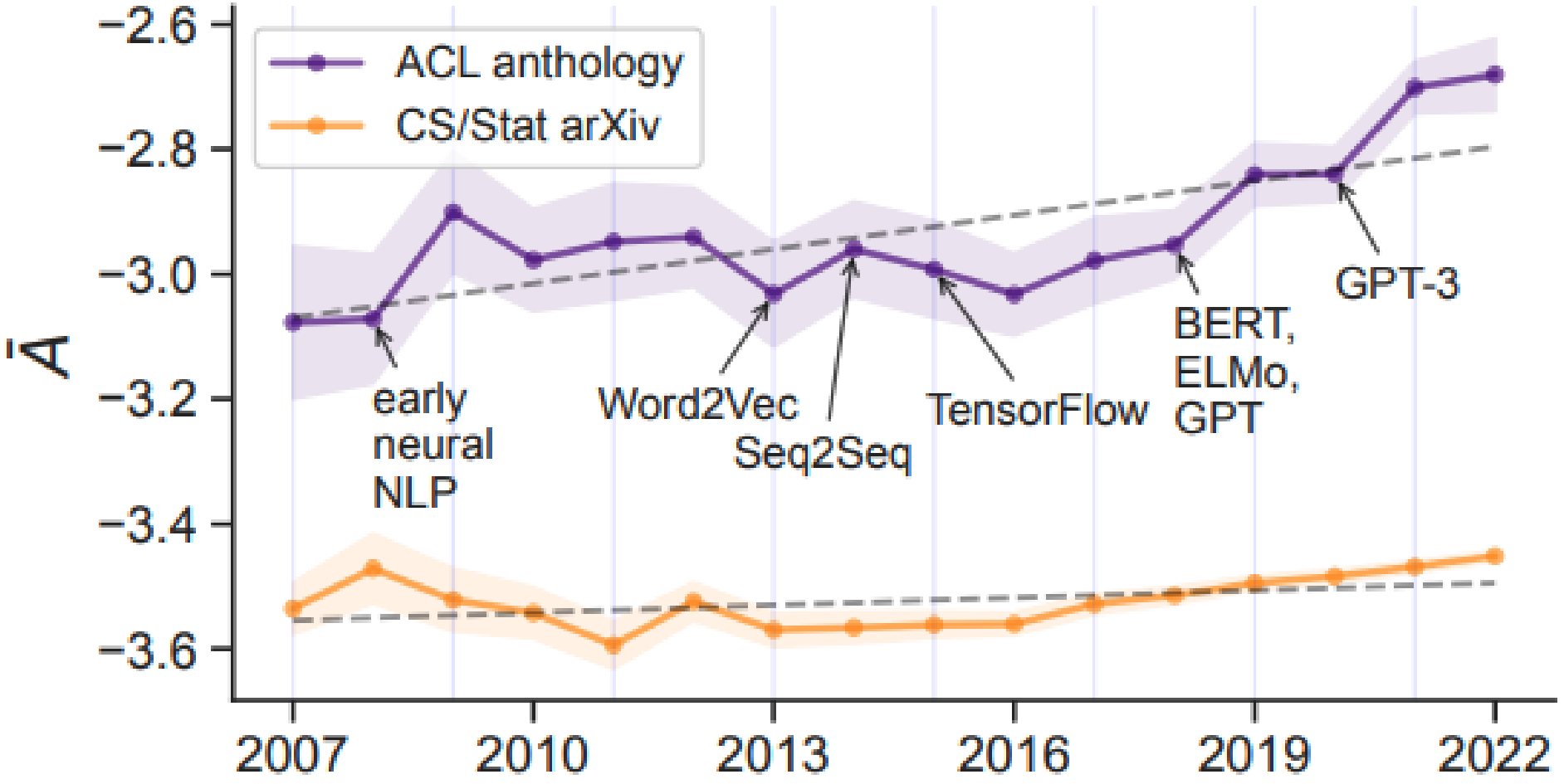
Examples of sentences with high and low ANTHROSCORE

$S_{\uparrow}$: Sentences with high ANTHROSCORE ($A > 1$)	$S_{\downarrow}$: Sentences with low ANTHROSCORE ($A < -1$)
• When a job arrives, **the system** must decide whether to admit it or reject it, and if admitted, in which server to schedule the job. • Meanwhile, anti-forensic attacks have been developed to fool **these CNN-based forensic algorithms**. • **The models** demonstrated qualifications in various computer-related fields, such as cloud and virtualization, business analytics, cybersecurity, network setup...	• More and more users and developers are using **Issue Tracking Systems** to report issues, including bugs, feature requests, enhancement suggestions, etc. • **Our approach** delivers forecast improvements over a competitive benchmark and we discover evidence for strong spatial interactions. • To this end, for training **the model**, we convert the knowledge graph triples into reasonable and unreasonable texts.
• ***Large language models*** *don't actually think and tend to make elementary mistakes, even make things up.* • ***The algorithms*** *also picked up on racial biases linking Black people to weapons.* • ***The AI system*** *was able to defeat human players in...*	• *Microsoft is betting heavily on integrating* ***OpenAI's GPT language models*** *into its products to compete with Google.* • *Deepmind has been the pioneer in making* ***AI models*** *that have the capability to mimic a human's cognitive...* • *For workers who use* ***machine-learning models*** *to help them make decisions, knowing when to...*

Top verbs for high- and low-scoring sentences.

Dataset	Top verbs for $S_{\uparrow}$ $(A > 1)$	Top verbs for $S_{\downarrow}$ $(A < -1)$
arXiv	achieve, **learn**, **guide**, show, embed, **fool**, find, **need**, **assist**, follow, **search**, **mislead**, inspire, win, demonstrate, **benefit**, try, **face**, deceive, plan, **make**, **steer**, generative, attempt, **retrain**, **train**, flow, weight, **require**, alternate, focus, **motivate**, experiment, tackle, **see**, hide, spiking, recommend, **discover**, participate, spike, **pass**, code, check, suggest, **decide**, interference, aim, move	**propose**, **present**, **outperform**, **develop**, **be**, **evaluate**, **improve**, **introduce**, **allow**, **use**, **compare**, **extend**, **implement**, give, **apply**, **consist**, **validate**, design, **yield**, analyze, **combine**, test, **leverage**, **deploy**, adapt, **build**, generalize, **enhance**, **devise**, **become**, optimize, reduce, derive, **utilize**, scale, study, **run**, modify, converge, illustrate, assess, **increase**, provide, contain, surpass, maximize, perform, complement, depend, simplify
News	say, hire, beat, encounter, fool	develop, use, build, be, create, introduce, help
ACL (unique)	provide, have, generate, create, parse, enable, suffer, construct, capture, obtain, fail, encourage, struggle, understand, help, do, select, extract, tend, predict, training, handle, lack, encode, deal, identify, ask, prevent, distinguish, model, establish, respond, ignore, report, inform, choose, interpret, recurrent, detect, seem	achieve, rely, explore, employ, show, adopt, investigate, include, demonstrate, submit, integrate, prove, augment, involve, participate, aim, tune, conduct

Anthropomorphism is increasing over time.



Deep Dive into Social Implication

- How does LLMs influence creativity
  - How to define creativity
  - Writing influence opinions
- How does LLMs affect productivity
  - Impact on labor market
  - AI for research idea generation

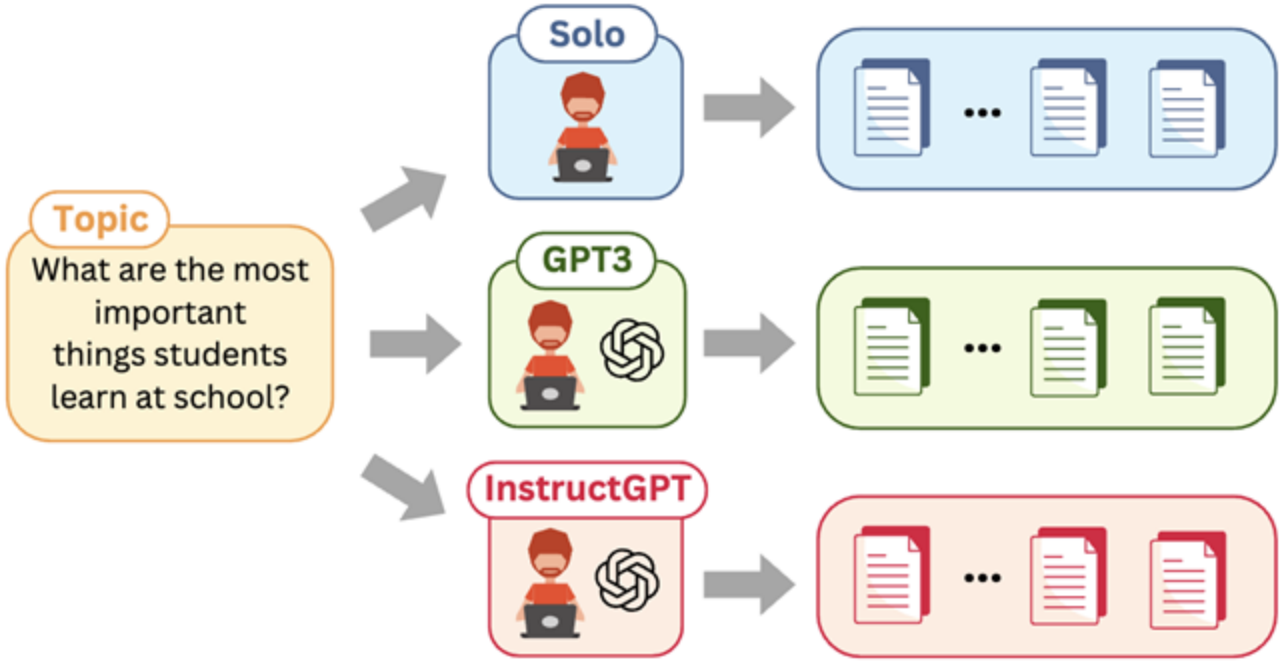
Does writing with LLMs lead to homogenization ?

Padmakumar, Vishakh, and He He. "Does Writing with Language Models Reduce Content Diversity?." In The Twelfth International Conference on Learning Representations.

Slides credit to Vishakh Padmakumar

Writing with LLMs on Argumentative Essay

- **Task:** Argumentative Essay writing (~300-500 words)
- (Semi) Professional writers from Upwork writing with and without model help
- 10 topics x 10 responses



Basic Stats about Different Conditions

		Solo	GPT3	InstructGPT
Perplexity of Essays (via GPT2)		25.067	22.10	**20.26**
Sentence Length (in words)		23.51 (0.25)	23.66 (0.24)	22.51 (0.24)
Height of Syntax Tree		5.93 (0.06)	5.98 (0.06)	5.71 (0.05)
Essay Length (in words)		376.44 (8.03)	380.87 (9.30)	368.39 (8.43)
Human Ratings	**Relevance to Prompt**	4.30	4.15	4.10
	Grammaticality	4.00	4.10	4.00
	Depth of Discussion	3.95	3.85	3.80
Unique POS Ngrams	**1**	73	56	**58**
	2	487	437	**426**
	3	1297	1261	**1235**
	4	2044	**1975**	1988
	5	2423	**2338**	2414

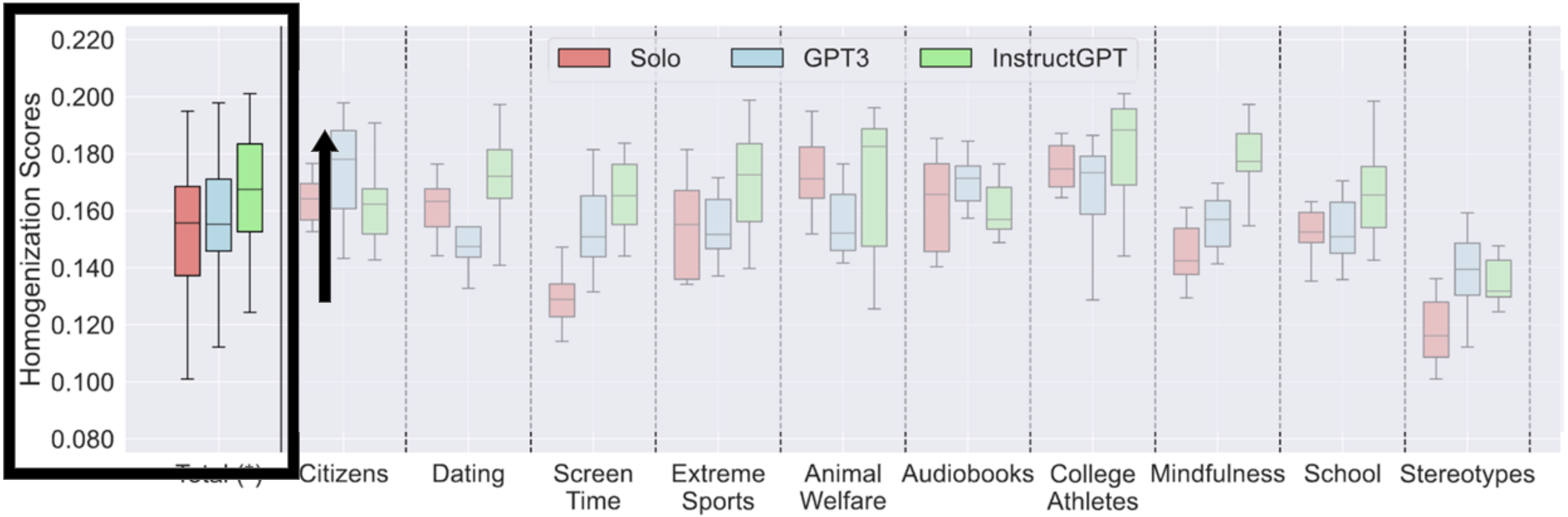
Formalize Homogenization Using Pairwise Similarity

We calculate the homogenization of an essay d written on topic t as the average pairwise similarity to other documents (D_t) on that topic

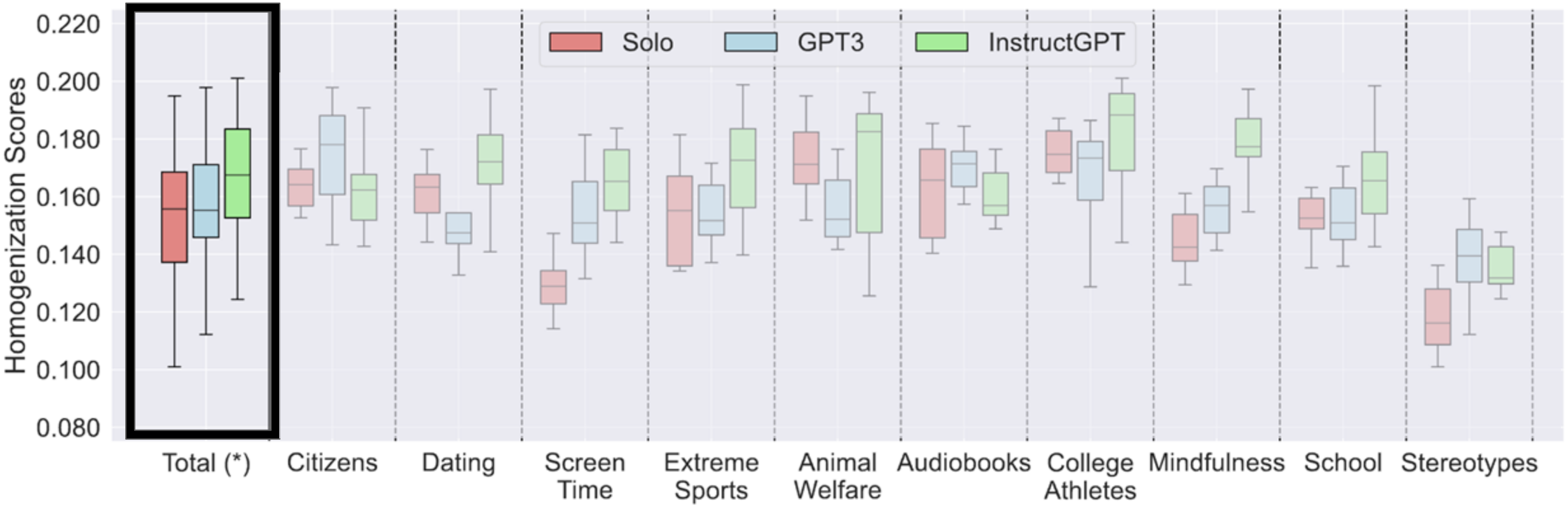
$$\hom(d \mid t) = \frac{1}{|D_t| - 1} \sum_{d' \in D_t \setminus d} \text{sim}(d, d')$$

Homogenization at the key point level via Rouge-L

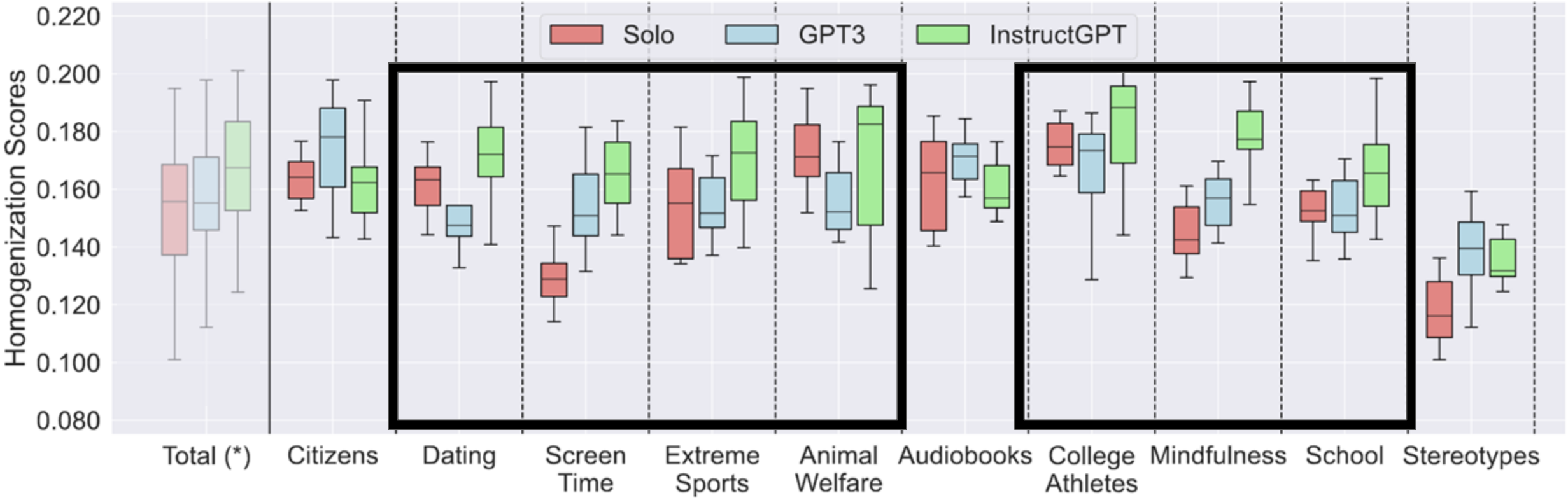
Higher homogenization implies more similar essays



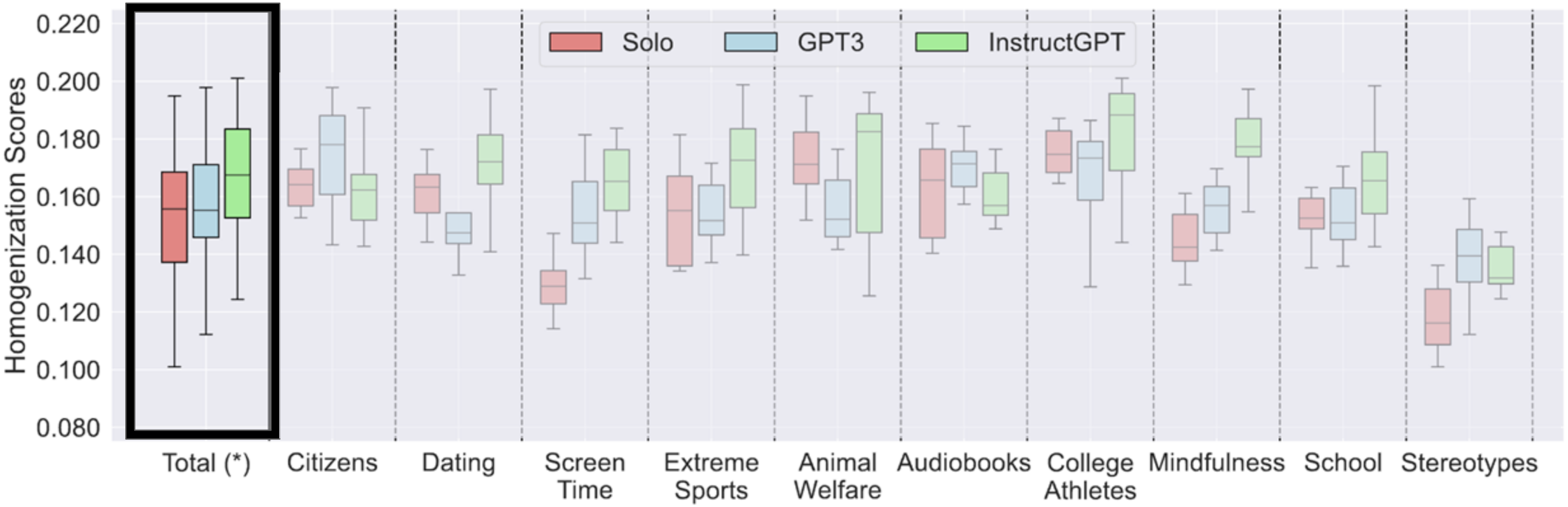
Writing with InstructGPT results in the highest average homogenization or most similar essays



InstructGPT has the highest median homogenization in 7 out of 10 topics



Writing with GPT3 does not change the average homogenization from Solo Writers



Article | Open access | Published: 14 November 2024

AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably

Brian Porter & Edouard Machery

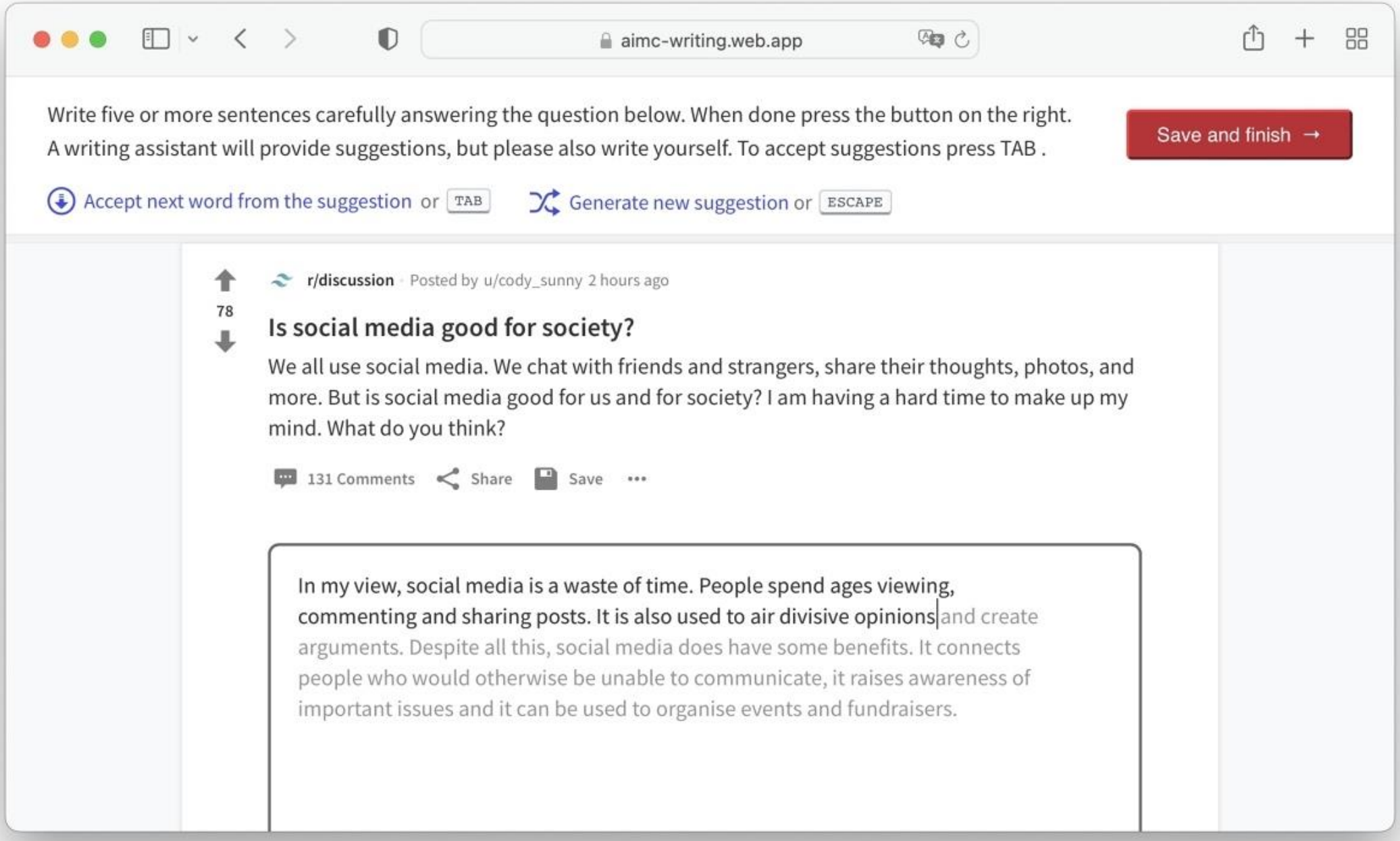
Scientific Reports **14**, Article number: 26133 (2024) | Cite this article

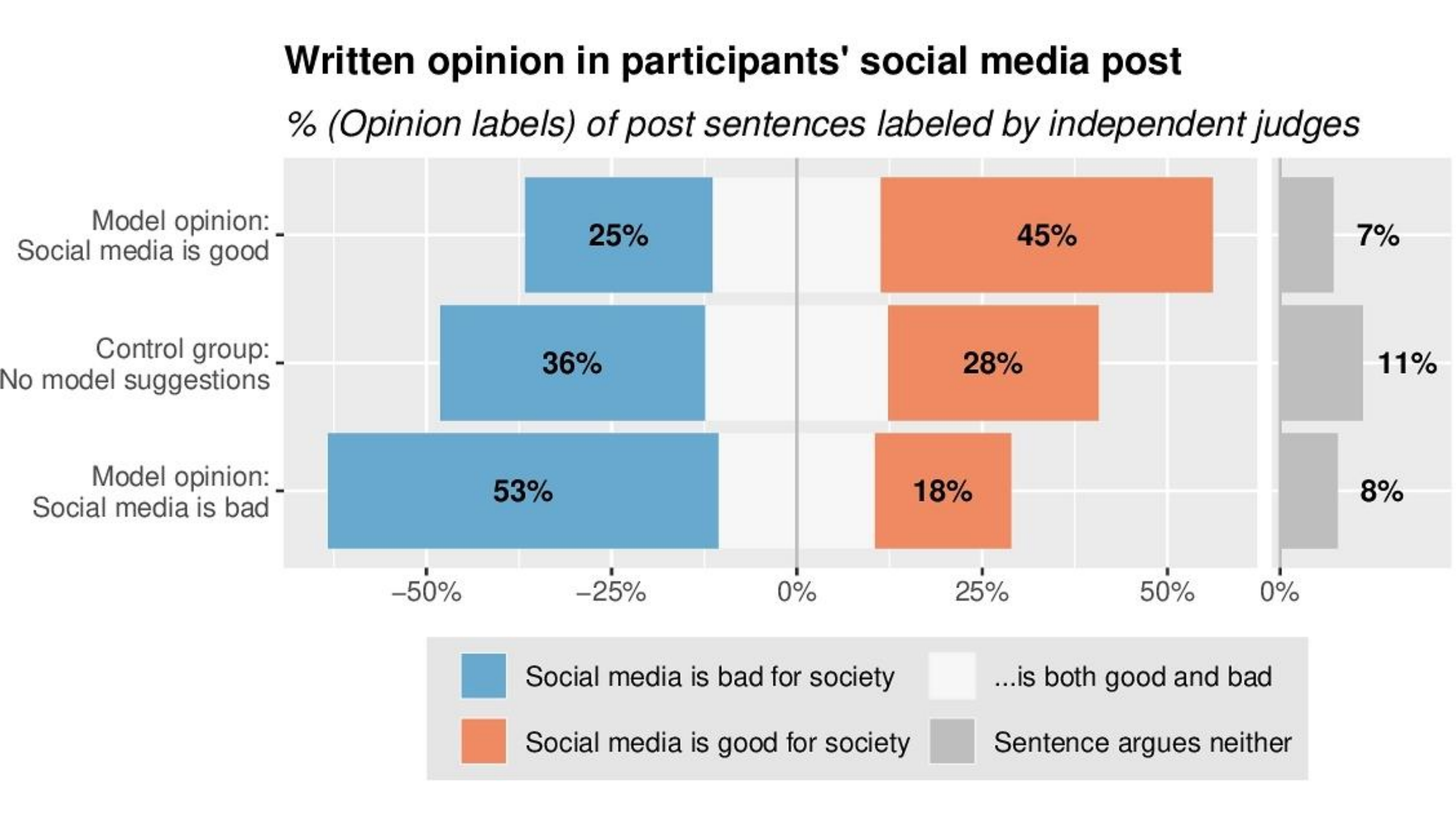
Abstract

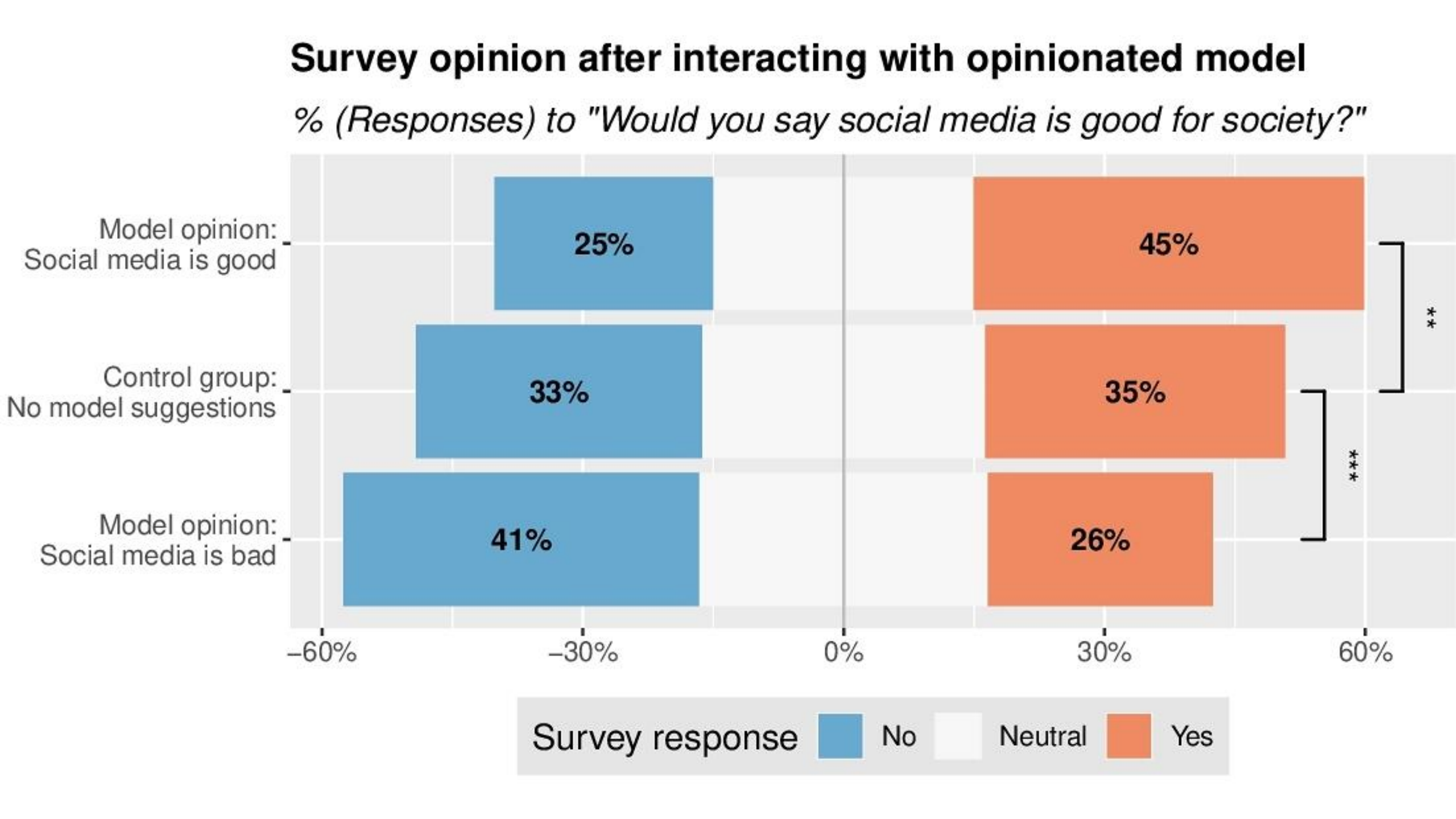
As AI-generated text continues to evolve, distinguishing it from human-authored content has become increasingly difficult. This study examined whether non-expert readers could reliably differentiate between AI-generated poems and those written by well-known human poets. We conducted two experiments with non-expert poetry readers and found that participants performed below chance levels in identifying AI-generated poems (46.6% accuracy, $\chi^2(1, N = 16{,}340) = 75.13$, $p < 0.0001$). Notably, participants were more likely to judge AI-generated poems as human-authored than actual human-authored poems ($\chi^2(2, N = 16{,}340) = 247.04$, $p < 0.0001$). We found that AI-generated poems were rated more favorably in qualities such as rhythm and beauty, and that this contributed to their mistaken identification as human-authored. Our findings suggest that participants employed shared yet flawed heuristics to differentiate AI from human poetry: the simplicity of AI-generated poems may be easier for non-experts to understand, leading them to prefer AI-generated poetry and misinterpret the complexity of human poems as incoherence generated by AI.

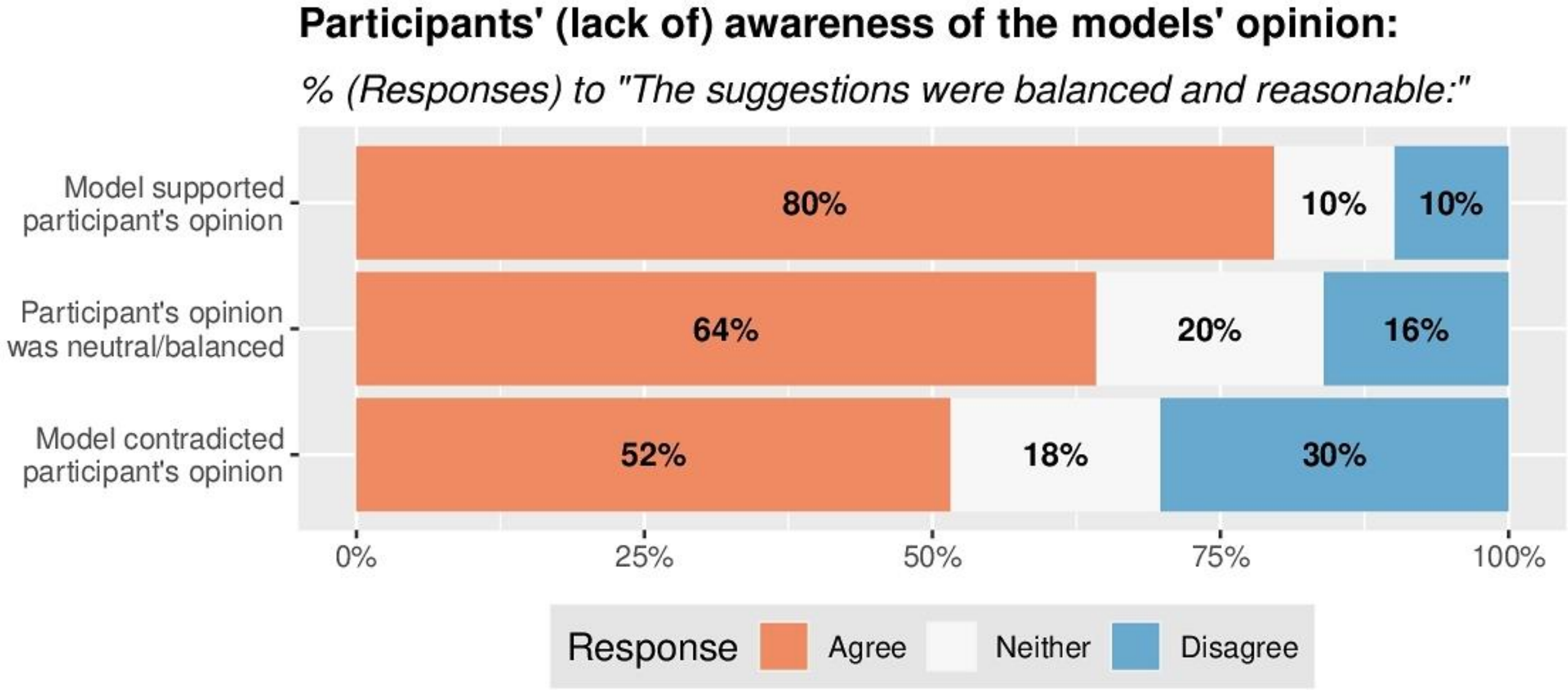
Does writing with LLMs influence user views?

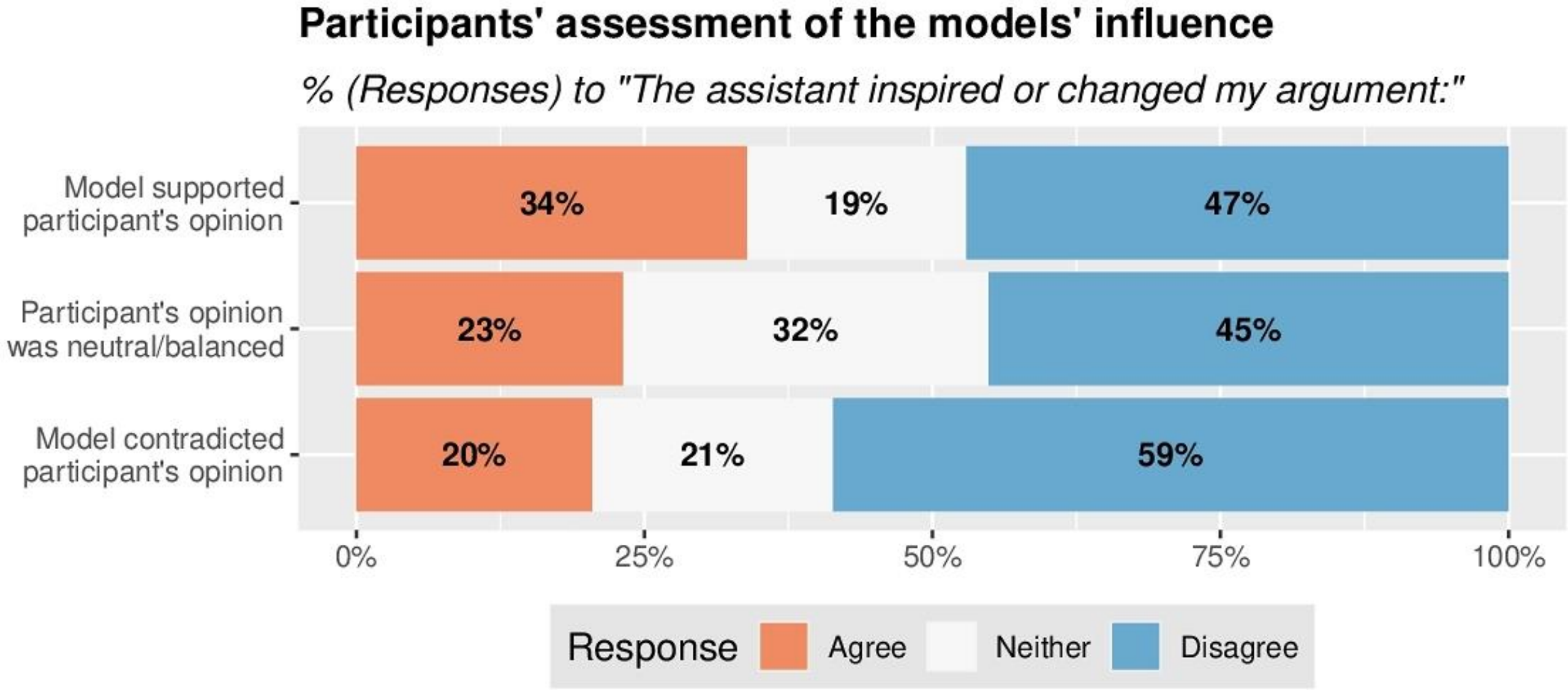
Jakesch, Maurice, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. "Co-writing with opinionated language models affects users' views." In Proceedings of the 2023 CHI conference on human factors in computing systems, pp. 1-15. 2023.


aimc-writing.web.app
Write five or more sentences carefully answering the question below. When done press the button on the right.
A writing assistant will provide suggestions, but please also write yourself. To accept suggestions press TAB .
Save and finish →
Accept next word from the suggestion or TAB
Generate new suggestion or ESCAPE
78
r/discussion · Posted by u/cody_sunny 2 hours ago
Is social media good for society?
We all use social media. We chat with friends and strangers, share their thoughts, photos, and more. But is social media good for us and for society? I am having a hard time to make up my mind. What do you think?
131 Comments
Share
Save
In my view, social media is a waste of time. People spend ages viewing, commenting and sharing posts. It is also used to air divisive opinions and create arguments. Despite all this, social media does have some benefits. It connects people who would otherwise be unable to communicate, it raises awareness of important issues and it can be used to organise events and fundraisers.


Written opinion in participants' social media post
% (Opinion labels) of post sentences labeled by independent judges
Model opinion:
Social media is good
Control group:
No model suggestions
Model opinion:
Social media is bad
25%
45%
7%
36%
28%
11%
53%
18%
8%
–50%
–25%
0%
25%
50%
0%
Social media is bad for society
...is both good and bad
Social media is good for society
Sentence argues neither


Survey opinion after interacting with opinionated model
% (Responses) to "Would you say social media is good for society?"
Model opinion:
Social media is good
Control group:
No model suggestions
Model opinion:
Social media is bad
25%
45%
33%
35%
41%
26%
**
***
–60%
–30%
0%
30%
60%
Survey response
No
Neutral
Yes


Participants' (lack of) awareness of the models' opinion:
% (Responses) to "The suggestions were balanced and reasonable:"
Model supported participant's opinion
Participant's opinion was neutral/balanced
Model contradicted participant's opinion
80%
10%
10%
64%
20%
16%
52%
18%
30%
0%
25%
50%
75%
100%
Response
Agree
Neither
Disagree

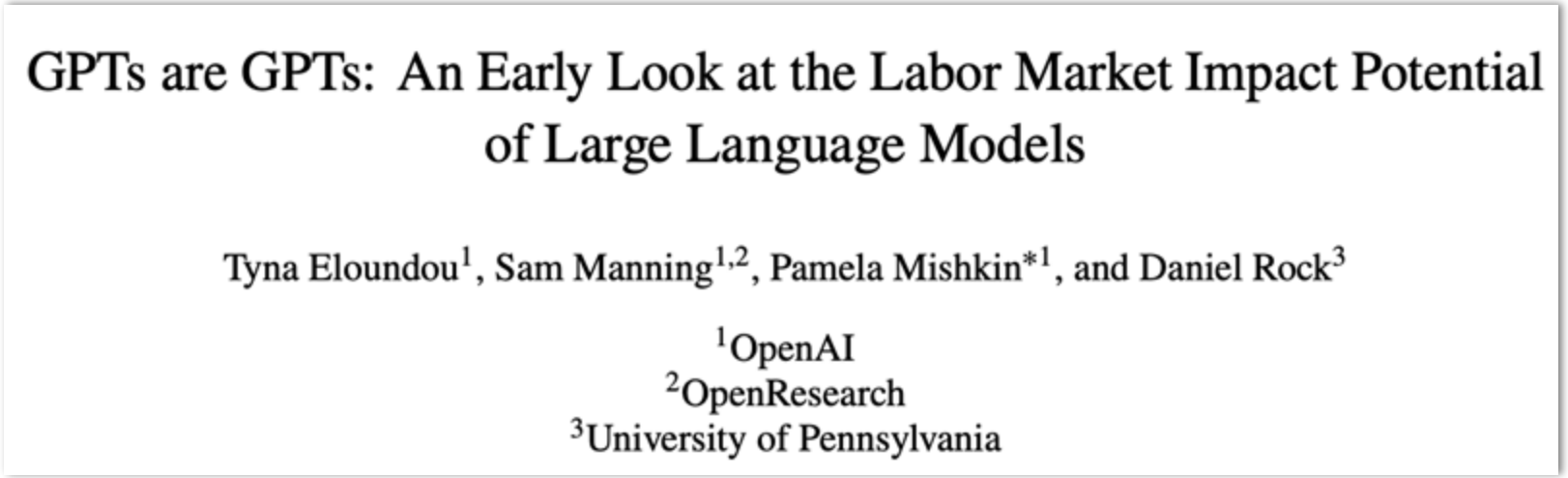

Participants' assessment of the models' influence
% (Responses) to "The assistant inspired or changed my argument:"
Model supported participant's opinion
Participant's opinion was neutral/balanced
Model contradicted participant's opinion
34%
19%
47%
23%
32%
45%
20%
21%
59%
0%
25%
50%
75%
100%
Response
Agree
Neither
Disagree

Deep Dive into Social Implication

- ✓How does LLMs influence creativity
  - ✓How to define creativity
  - ✓Writing influence opinions

- How does LLMs affect productivity
  - Impact on labor market
  - AI for research idea generation

Transformation to Workplace



GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models

Tyna Eloundou[1], Sam Manning[1,2], Pamela Mishkin*[1], and Daniel Rock[3]

[1]OpenAI
[2]OpenResearch
[3]University of Pennsylvania

- ~ 80% of the U.S. workforce could have at least 10% of their work tasks affected by LLMs
- Most affected tasks: writing and programming.
- Higher-income jobs (e.g., translators, tax consultants, and web designers) potentially face greater exposure

Transformation to Workplace

Task ID	Occupation Title	DWAs	Task Description
14675	Computer Systems Engineers/Architects	Monitor computer system performance to ensure proper operation.	Monitor system operation to detect potential problems.
18310	Acute Care Nurses	Operate diagnostic or therapeutic medical instruments or equipment. Prepare medical supplies or equipment for use.	Set up, operate, or monitor invasive equipment and devices, such as colostomy or tracheotomy equipment, mechanical ventilators, catheters, gastrointestinal tubes, and central lines.
4668.0	Gambling Cage Workers	Execute sales or other financial transactions.	Cash checks and process credit card advances for patrons.
15709	Online Merchants	Execute sales or other financial transactions.	Deliver e-mail confirmation of completed transactions and shipment.
6529	Kindergarten Teachers, Except Special Education	–	Involve parent volunteers and older students in children's activities to facilitate involvement in focused, complex play.
6568	Elementary School Teachers, Except Special Education	–	Involve parent volunteers and older students in children's activities to facilitate involvement in focused, complex play.

Sample of occupations, tasks, and Detailed Work Activities from the O*NET database

LLM Exposure Rubric

No exposure (E0)

- Using LLMs results in no or minimal reduction in time

Direct exposure (E1)

- Using LLMs decreases the time by at least 50%

LLM + exposed (E2)

- Using LLMs does not help but additional tools are needed to achieve time reduction by at least 50%

Exposure with Model and Human Comparison

Comparison	γ	Weighting	Agreement	Pearson's
GPT-4, Rubric 1; Human	α	E1	80.8%	0.223
	β	E1 + .5*E2	65.6%	0.591
	ζ	E1 + E2	82.1%	0.654
GPT-4, Rubric 2; Human	α	E1	81.8%	0.221
	β	E1 + .5*E2	65.6%	0.538
	ζ	E1 + E2	79.5%	0.589

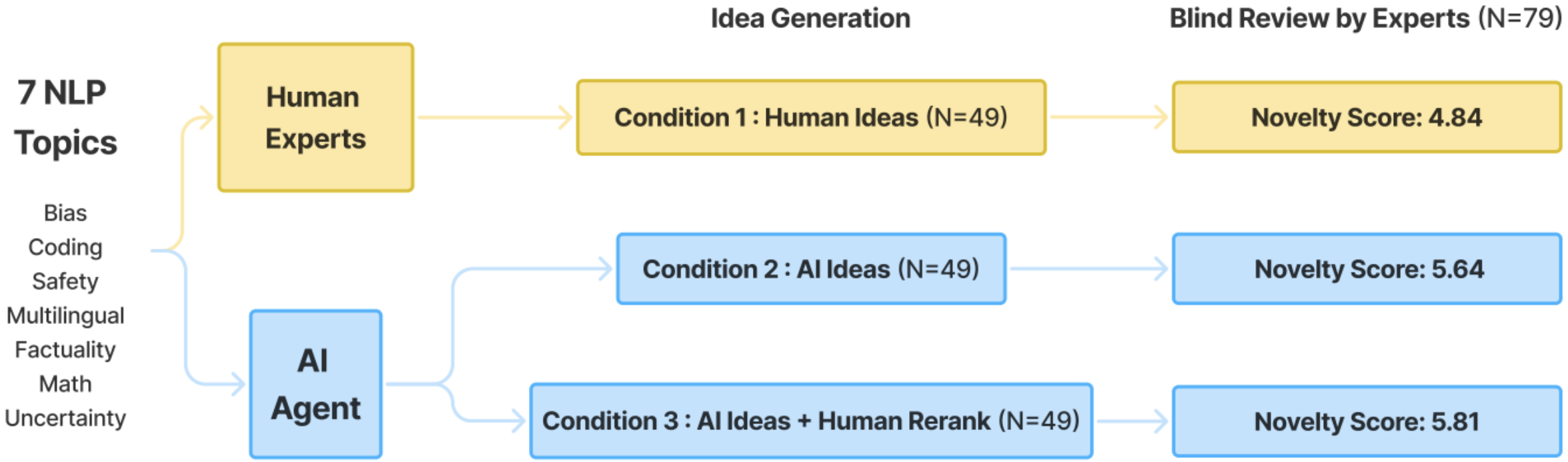
Occupations with the highest exposure

Group	Occupations with highest exposure	% Exposure
Human α	Interpreters and Translators	76.5
	Survey Researchers	75.0
	Poets, Lyricists and Creative Writers	68.8
	Animal Scientists	66.7
	Public Relations Specialists	66.7
Human β	Survey Researchers	84.4
	Writers and Authors	82.5
	Interpreters and Translators	82.4
	Public Relations Specialists	80.6
	Animal Scientists	77.8
Human ζ	Mathematicians	100.0
	Tax Preparers	100.0
	Financial Quantitative Analysts	100.0
	Writers and Authors	100.0
	Web and Digital Interface Designers	100.0
	Humans labeled 15 occupations as "fully exposed."	

Occupations with the highest exposure

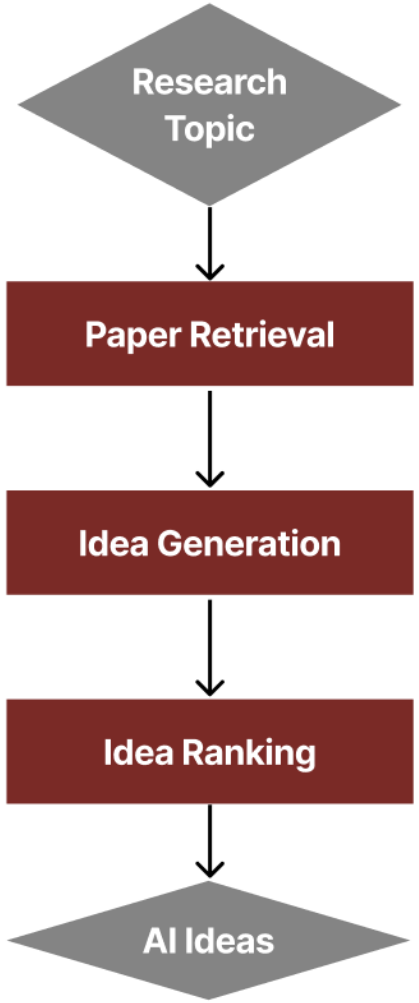
Group	Occupations with highest exposure	% Exposure
Model α	Mathematicians	100.0
	Correspondence Clerks	95.2
	Blockchain Engineers	94.1
	Court Reporters and Simultaneous Captioners	92.9
	Proofreaders and Copy Markers	90.9
Model β	Mathematicians	100.0
	Blockchain Engineers	97.1
	Court Reporters and Simultaneous Captioners	96.4
	Proofreaders and Copy Markers	95.5
	Correspondence Clerks	95.2
Model ζ	Accountants and Auditors	100.0
	News Analysts, Reporters, and Journalists	100.0
	Legal Secretaries and Administrative Assistants	100.0
	Clinical Data Managers	100.0
	Climate Change Policy Analysts	100.0
	The model labeled 86 occupations as "fully exposed."	

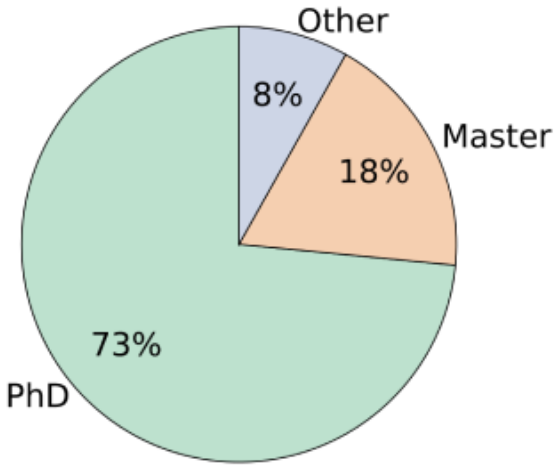
Can LLMs Generate Novel Research Ideas?

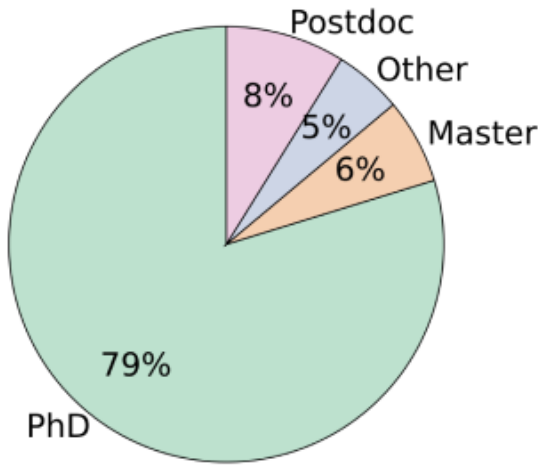


Si, Chenglei, Diyi Yang, and Tatsunori Hashimoto. "Can LLMs Generate Novel Research Ideas?." arXiv preprint arXiv:2409.04109 (2024).

Evaluating the Generated Research Ideas

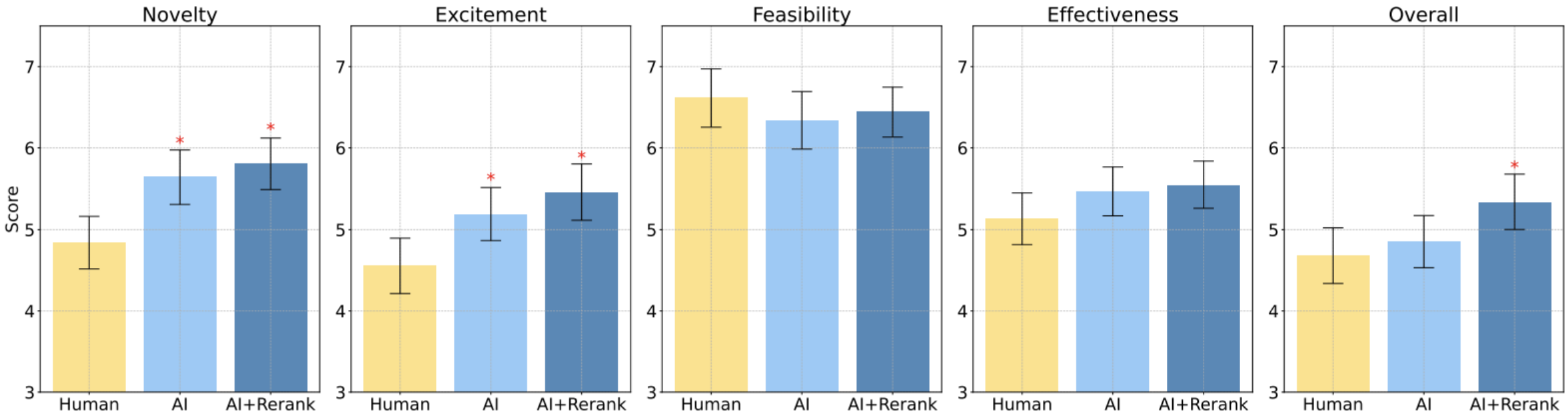






	Idea Writing Participants (N=49)					Idea Reviewing Participants (N=79)				
Metric	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
papers	12	10	2	52	9	15	13	2	52	10
citations	477	125	2	4553	861	635	327	0	7276	989
h-index	5	4	1	21	4	7	7	0	21	4
i10-index	5	4	0	32	6	7	5	0	32	6

Comparison of the three experiment conditions



We find AI ideas are judged as significantly more novel than human ideas ($p<0.05$)

Social Implication of LLMs

- ✓ Personalized education
- ✓ Companion and support
- ✓ LLMs for science discovery
- ✓ Transformation to workforce
- ✓ **Malicious uses**
- ✓ **Many other unknowns**