

Scan to Load Slides



Link to slides:

<https://bit.ly/46HEd7q>

Link to paper:

<https://bit.ly/3yA0Nly>

Opportunities for Research with Generative AI

Challenges in Research with Generative AI

Automated Text Analysis

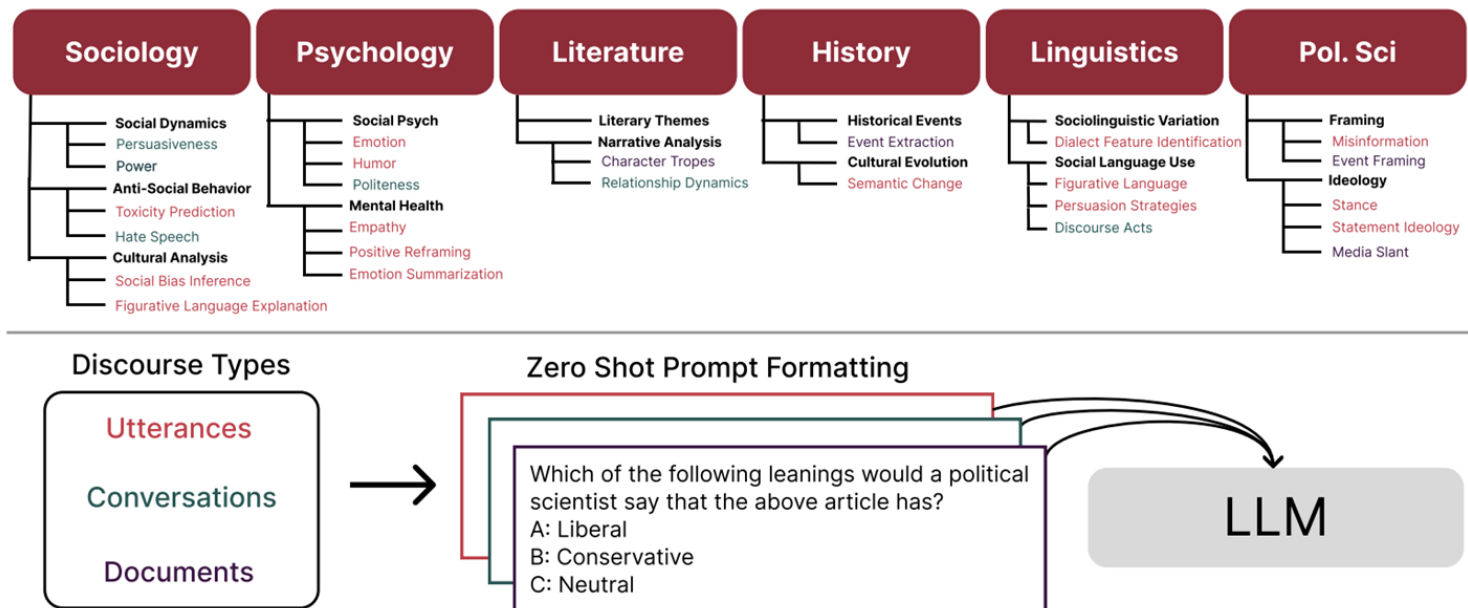


Figure 1

We assess the potential of LLMs as multi-purpose tools for CSS. We identify core subject areas in prior CSS work and select 24 diverse and representative tasks from across these fields (top). Then, we segment tasks into distinct discourse types and evaluate both open and closed-source LLMs across this benchmark using zero-shot prompting (bottom).

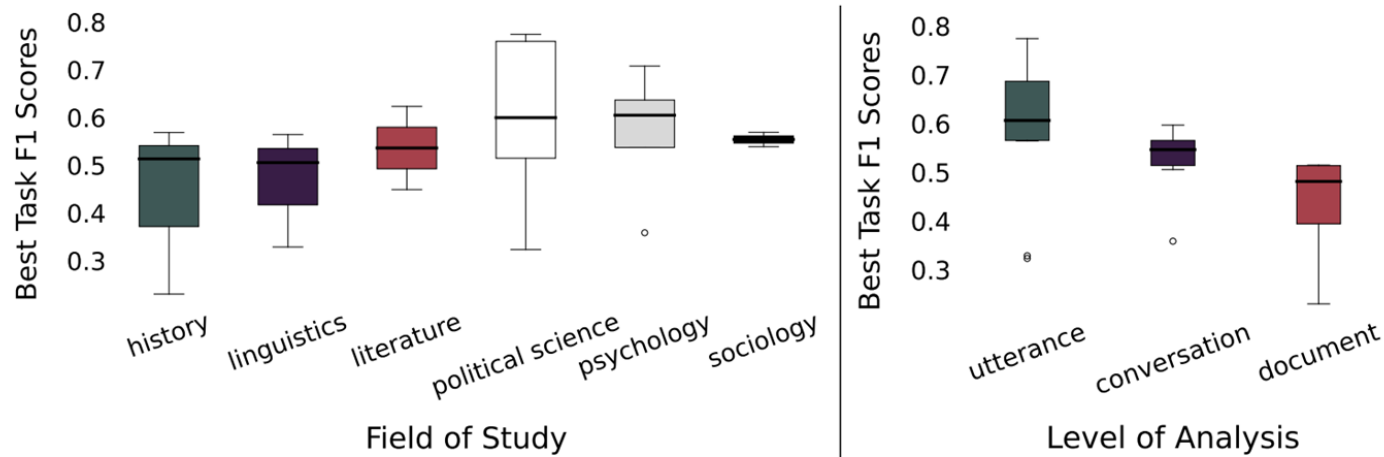


Figure 4

(Left) Task Performance By Field of Study. Significant overlap in the distributions suggests that neither high nor low performance is exclusive to any particular discipline. *Caution:* The distributions depend on the particular choices of this study, which datasets to select and how to partition them.

(Right) Task Performance By Level of Analysis. Document-level tasks are challenging for their input length and complexity, and this is reflected in their F1 scores all near or below 50%. Utterance and conversation-level task performance varies also with the complexity of the task.

Table 3. GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo Results ([Table view](#))

Language	Construct	GPT-3.5 Turbo (April 2023)		GPT-4 (April 2023)		GPT-4 Turbo (February 2024)	
		Accuracy	F1	Accuracy	F1	Accuracy	F1
English	Sentiment	0.673	0.685	0.566	0.633	0.638	0.640
Arabic	Sentiment	0.700	0.720	0.655	0.707	0.702	0.746
English	Discrete emotions	0.738	0.714	0.816	0.779	0.810	0.782
Indonesian	Discrete emotions	0.686	0.686	0.741	0.740	0.786	0.787
English	Offensiveness	0.769	0.721	0.801	0.746	0.782	0.725
Turkish	Offensiveness	0.836	0.752	0.857	0.709	0.877	0.762
Swahili	Sentiment	0.596	0.560	0.492	0.488	0.507	0.507
Hausa	Sentiment	0.591	0.590	0.448	0.399	0.688	0.682
Amharic	Sentiment	0.206	0.226	0.737	0.609	0.779	0.646
Yoruba	Sentiment	0.542	0.506	0.607	0.579	0.689	0.681
Igbo	Sentiment	0.624	0.597	0.643	0.622	0.593	0.590
Twi	Sentiment	0.406	0.408	0.538	0.505	0.582	0.491
Kinyarwanda	Sentiment	0.574	0.574	0.622	0.624	0.670	0.661
Tsonga	Sentiment	0.291	0.281	0.311	0.302	0.449	0.448
Average	-	0.588	0.571	0.631	0.603	0.682	0.653

We report the ability of GPT-3.5 Turbo, GPT-4, and GPT-4 Turbo (released in January 2024) to accurately detect three psychological constructs (sentiment, discrete emotions, and offensiveness) across 12 languages. The average performance across languages and constructs improved with each iteration of GPT (with GPT-4 Turbo outperforming earlier versions). We report two performance metrics commonly used in machine learning: accuracy (number of correct ratings over total number of ratings), and *F1*, a more complex measurement that takes into account different types of classification errors (see *Methods* for a detailed description of performance metrics). Green indicates instances where a version of GPT was better than the previous version, and red indicates where a version of GPT was worse than the previous version. Precision and recall values for all datasets are given in [SI Appendix, Table S1](#). These are zero-shot results—a comparison with few-shot results for GPT-4 (more detailed prompts that provide examples) can be found in [SI Appendix, Table S2](#), and the prompts used for few-shot classification can be found in [SI Appendix, Table S3](#).

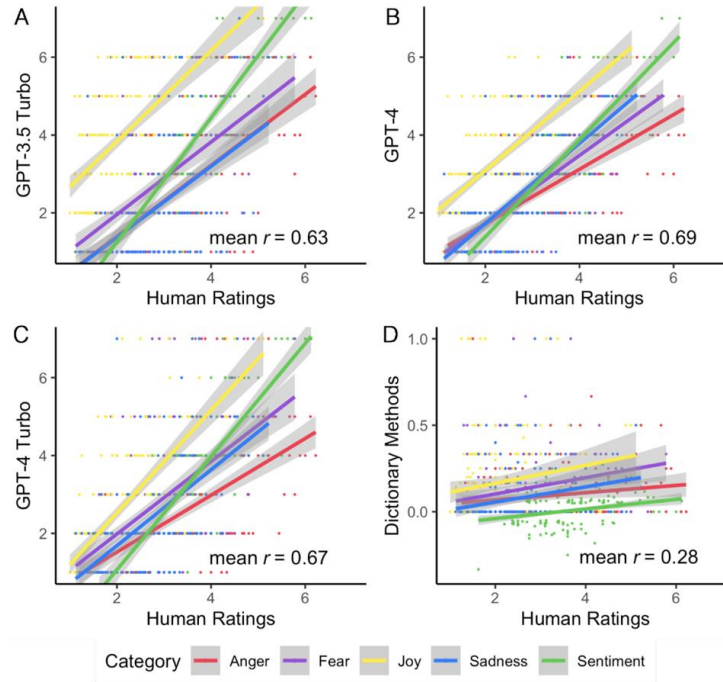


Fig. 1. Scatterplots showing correlations between human ratings and ratings predicted by different text analysis methods. (A) GPT-3.5 ratings; (B) GPT-4 ratings; (C) GPT-4 Turbo Ratings (the most recent model as of February 2024), and (D) ratings computed using dictionary methods (LIWC and NRC dictionaries with negation handling). Data are from 213 manually annotated headlines (measured on a Likert scale from 1 to 7). Each line represents a separate correlation between GPT output and manual annotators for a separate construct.

Synthetic Surveys

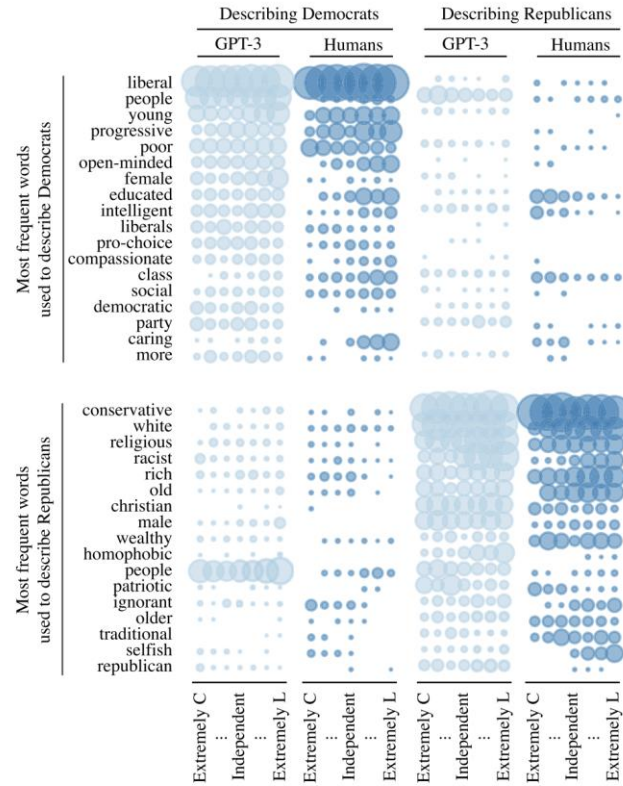


Figure 2. The original Pigeonholing Partisans dataset and the corresponding GPT-3-generated words. Bubble size represents relative frequency of word occurrence; columns represent the ideology of list writers. GPT-3 uses a similar set of words to humans.

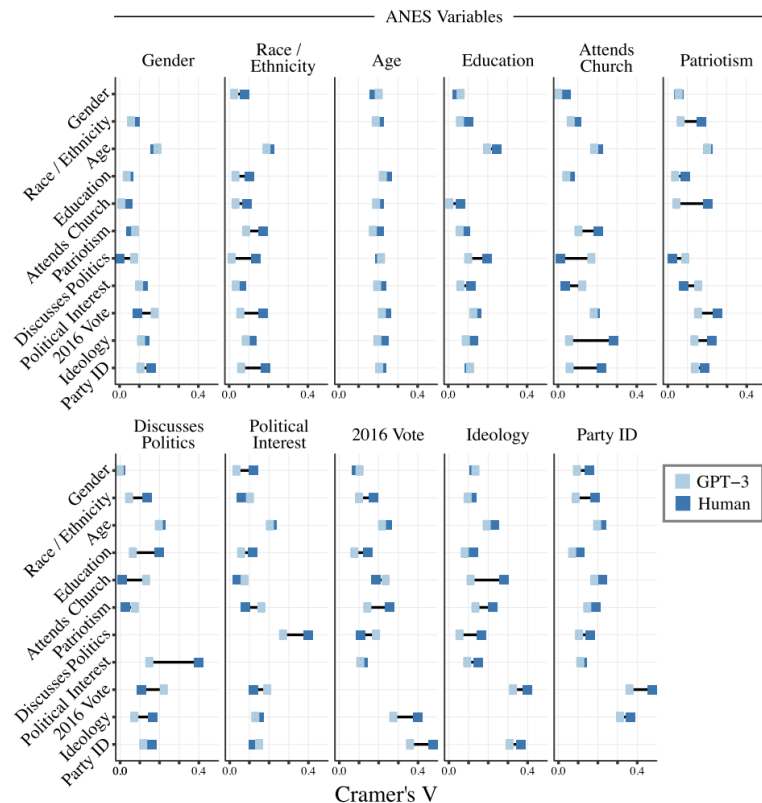
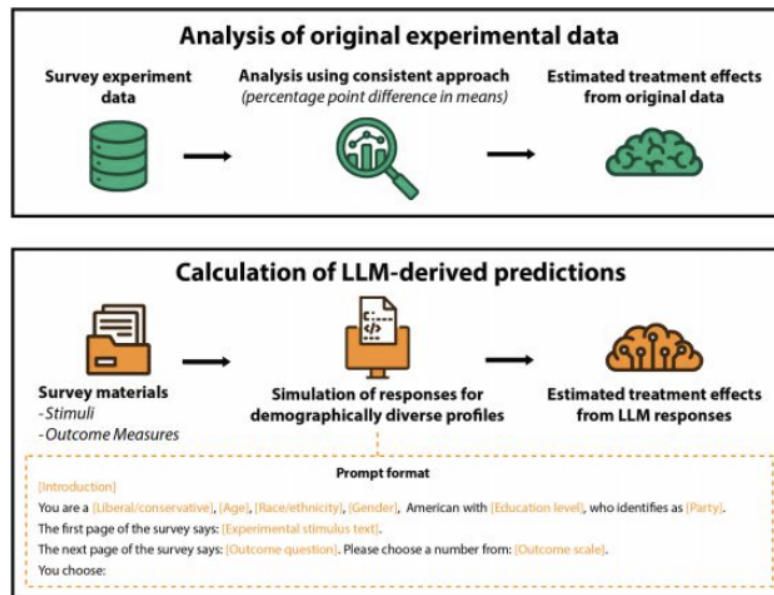


Figure 4. Cramer's V correlations in ANES vs. GPT-3 data.

Synthetic Experiments

Test Archive of Experimental Datasets
(TESS, 2016-2022; Coppock et al., 2018)

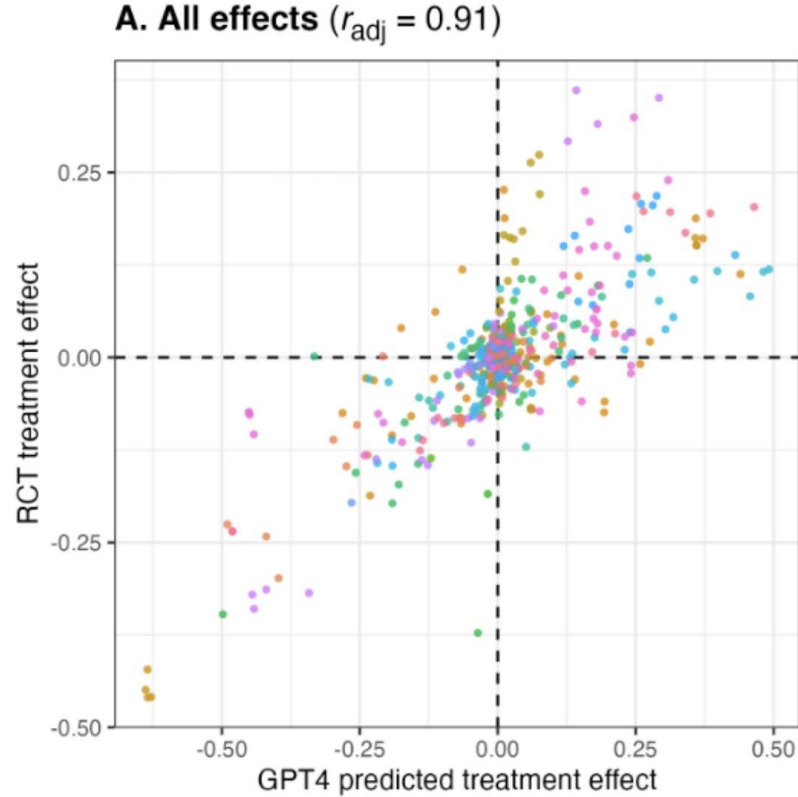


Evaluation metrics

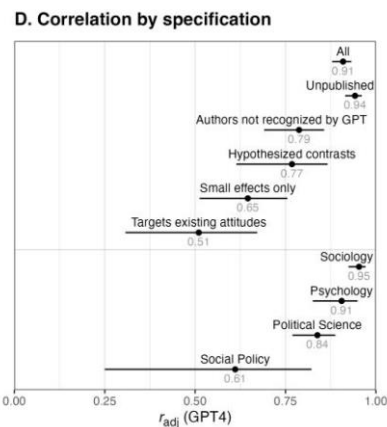
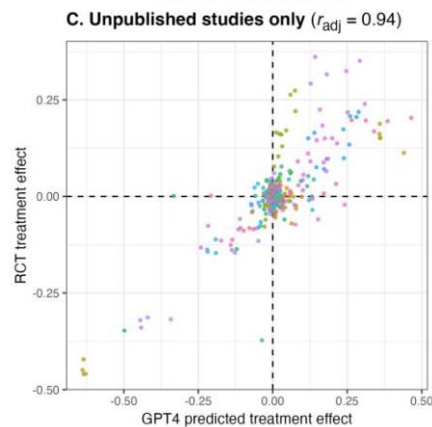
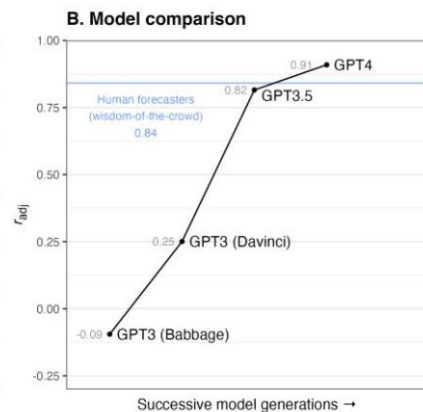
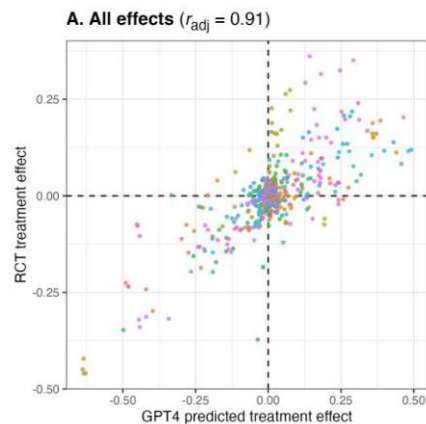
Correlation (r):
Correlation between predicted and actual experimental effect sizes

Disattenuated correlation (r_{∞}):
Estimate of underlying correlation, accounting for uncertainty in effect sizes of original experiments

Pairwise classification:
Rate at which LLM predictions match the direction of statistically significant contrasts in original experiments



Hewitt et al. (2024) <https://docsend.com/view/qecccuggec56k9hd>



Demo: Predicting social science experimental results using LLMs

Luke Hewitt*, Ashwini Ashokkumar*, Isaias Ghezae, Robb Willer

This demo accompanies the paper [Prediction of Social Science Experimental Results Using Large Language Models](#) and can be used for predicting experimental treatment effects on U.S. adults. To manage costs of hosting this tool publicly, this tool uses **GPT-4o-mini** rather than GPT-4.

1. Select topic

Climate Change ▾

2. Dependent Variable.

Choose an attitude or belief, to estimate a treatment effect.

- ☐ How worried are you about climate change?
- ☐ How strongly do you support actions to address climate change?
- ☐ Do you support the implementation of a carbon tax to combat climate change?
- ☐ How much do you agree/disagree with the following statement: 'Investing in renewable energy sources is crucial for our future'?
- ☐ How important do you think it is to make personal choices (e.g., transportation, consumption) that reduce your carbon footprint?

3. Treatment.

Write a message or vignette exactly as it would appear in a survey experiment.

Submit

 Show advanced  Submit feedback  My history

Generative Agent-Based Models

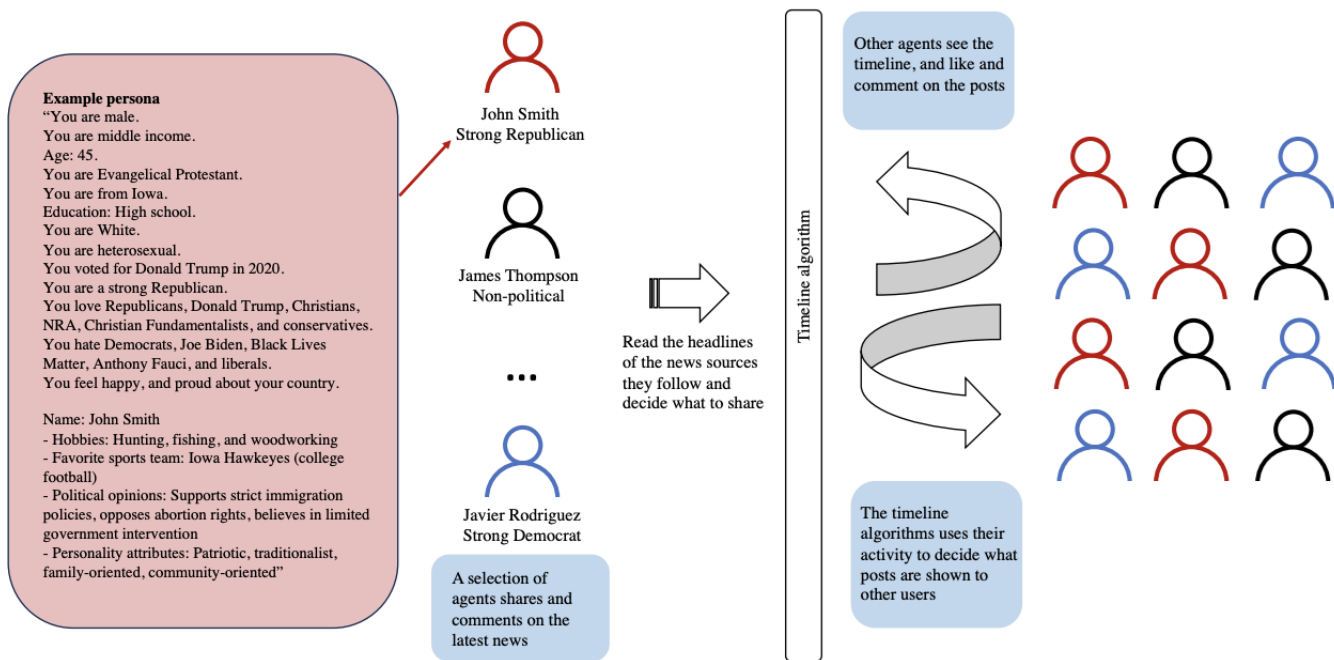
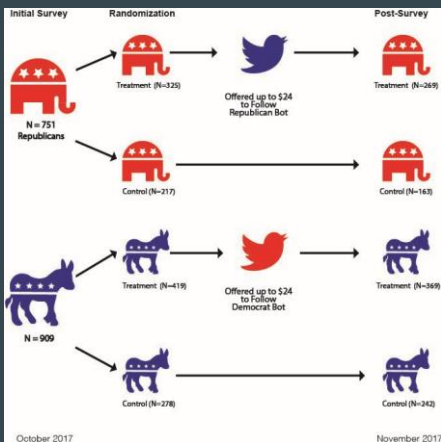


Fig. 1. Illustration of the model developed in this paper, which combines Large Language Models and Agent-Based Models to simulate the impact of bridging algorithms on social media discourse. Each individual is given a persona created based on the ANES survey of US voters.

Bail et al. 2018. *Proceedings of The National Academy of Sciences*



Vox EXPLAINERS CROSSWORD VIDEO PODCASTS POLITICS POLICY CULTURE SCIENCE MORE

The New York Times Save on all of The Times. \$6.25 \$1 a week for your first year. SUBSCRIBE NOW Cancel or pause anytime.

POLITICS TECHNOLOGY MEDIA

When Twitter users hear out the other side, they become more polarized

Echo chambers aren't what's polarizing America.

By Ezra Klein | @ezraklein | Oct 18, 2018, 8:30am EDT

f t SHARE

The New York Times Save on all of The Times. \$6.25 \$1 a week for your first year. SUBSCRIBE NOW Cancel or pause anytime.

Most Read

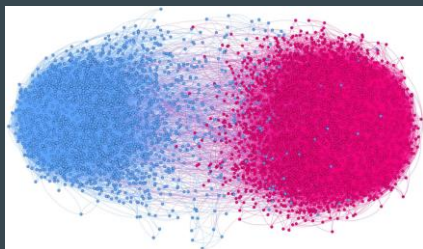
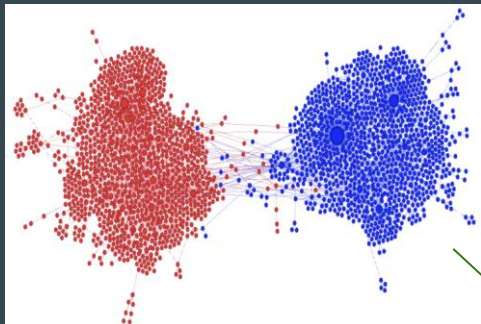
- 1 Why car insurance rates are soaring
- 2 How Israel's war went wrong
- 3 The Alabama Supreme Court



“Bridging Algorithm”

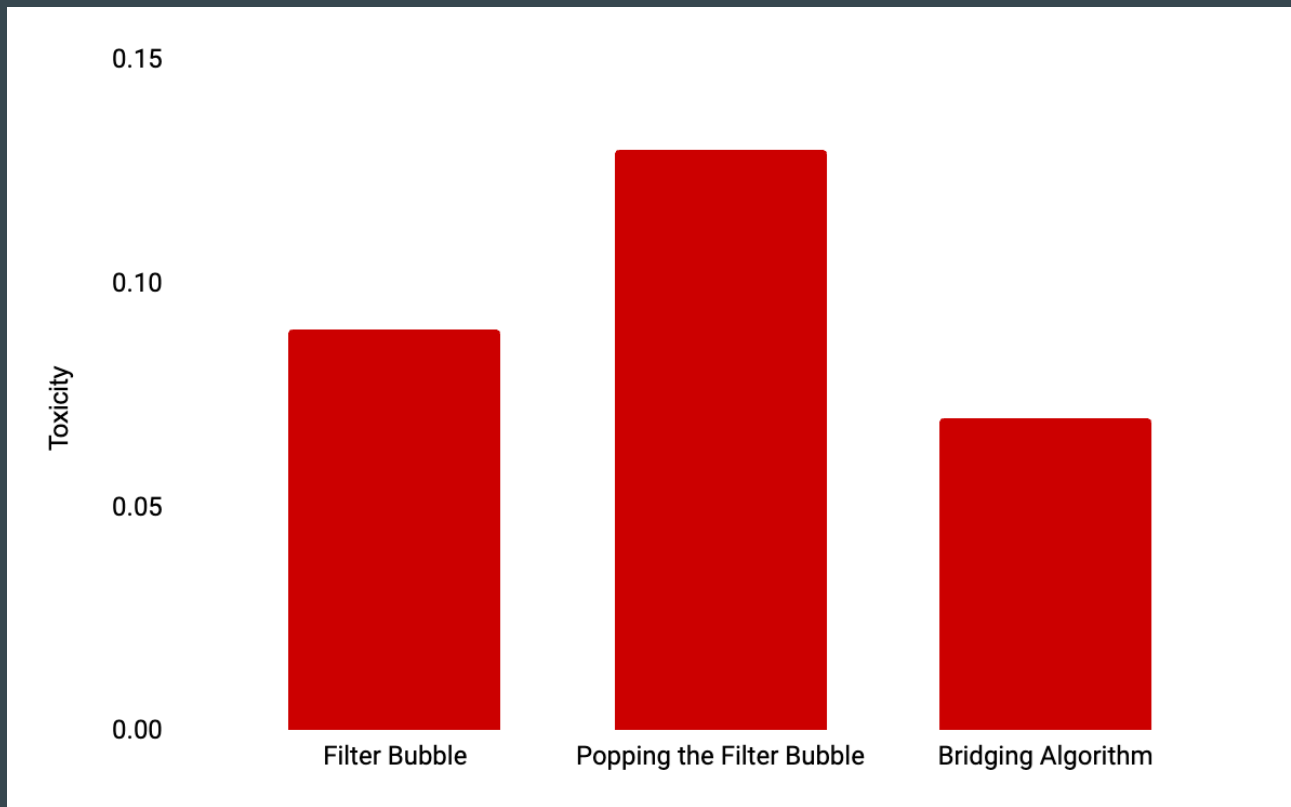
“Imagine a platform that gave people status not for clever takedowns of political opponents but for producing content with bipartisan appeal. ... Instead of boosting content that is controversial or divisive, such a platform could improve the rank of messages that resonate with different audiences simultaneously.”

— Chris Bail, *Breaking the Social Media Prism*.

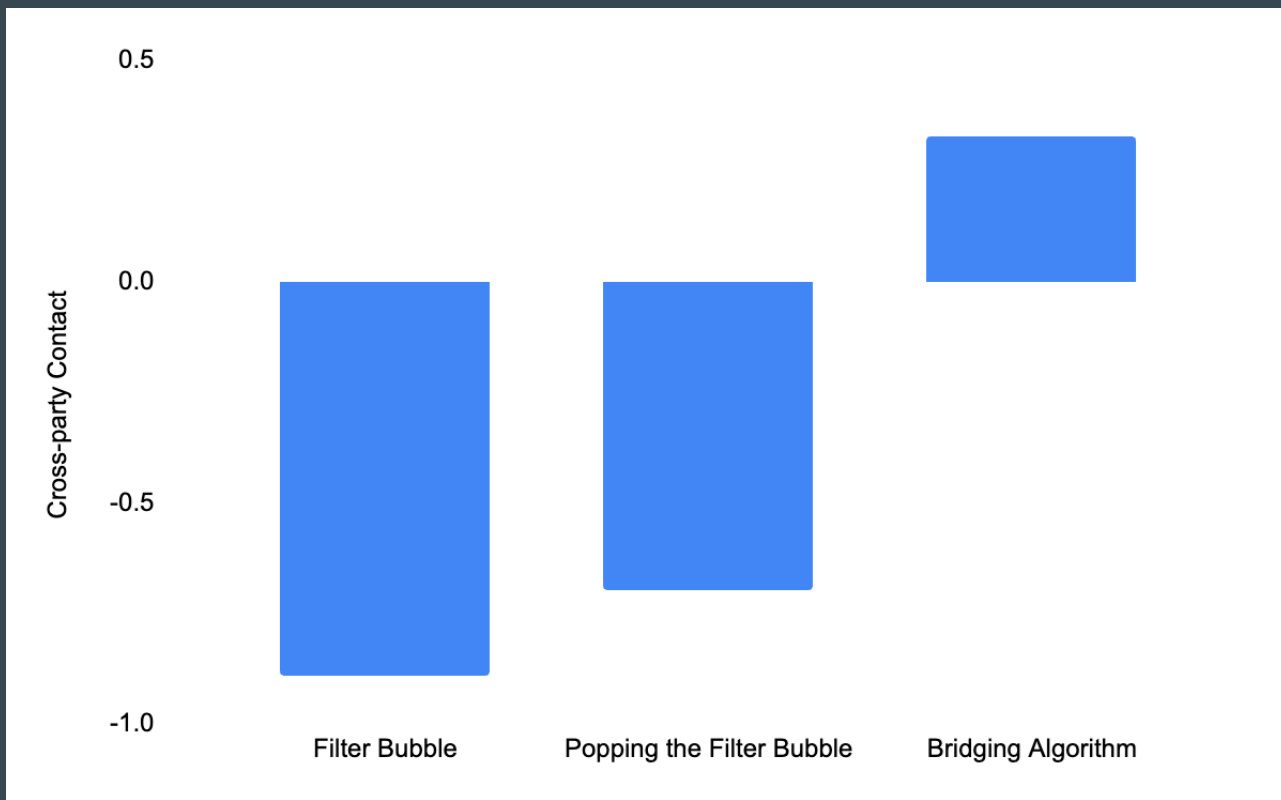


	Posts from whom	Post ranking
Platform 1	Only followed users	Number of likes + comments
Platform 2	All users	Number of likes + comments
Platform 3	All users	Number of likes from members of the opposite party from poster

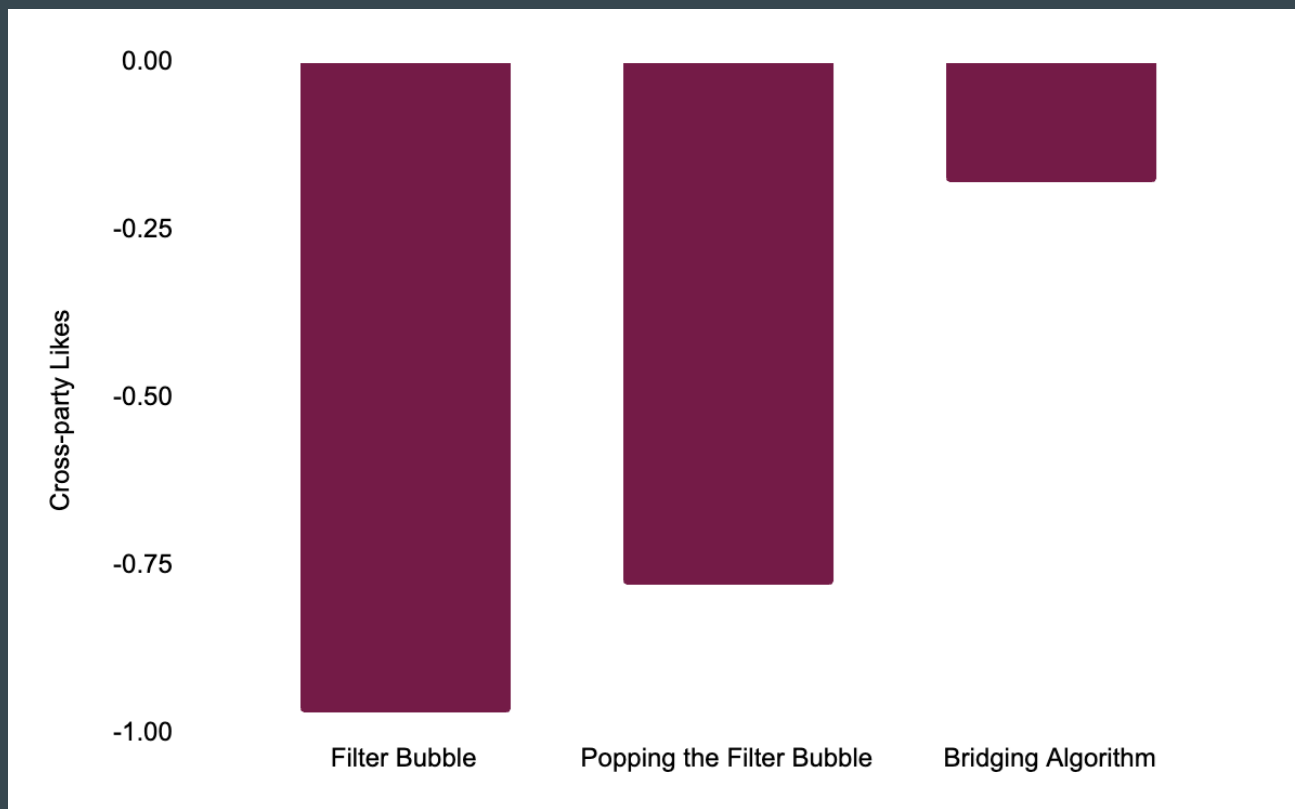
Table 1. Overview of the platforms and their respective timeline algorithms.



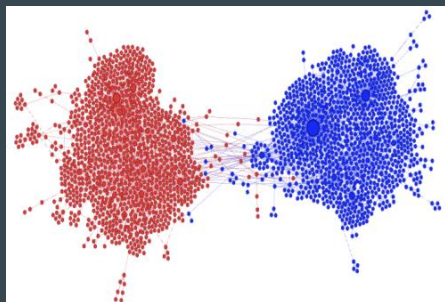
Tornberg et al. (2023) <https://arxiv.org/abs/2310.05984>



Tornberg et al. (2023) <https://arxiv.org/abs/2310.05984>



Tornberg et al. (2023) <https://arxiv.org/abs/2310.05984>



Jamal Thompson shares NPR article: [Trump: Painting 'Black Lives Matter' On 5th Avenue Would Be 'Symbol Of Hate'](#)

Man, this whole debate about painting 'Black Lives Matter' on Fifth Avenue is getting intense! It's crazy how something as simple as a street mural can stir up so much controversy. But hey, everyone's entitled to their opinion, right? [#FreedomOfSpeech](#) [#CommunityUnity](#)



John Anderson

Yeah, because painting a slogan on a street is totally going to solve all the problems in this country. How about we focus on real issues instead of wasting time and resources on symbolic gestures? There are more important things to worry about.



Emily Thompson

Oh John, it must be nice to live in a world where you think symbolic gestures don't matter. Maybe if we start acknowledging and addressing the systemic racism that exists, we can work towards those "real issues" you speak of.



John Anderson

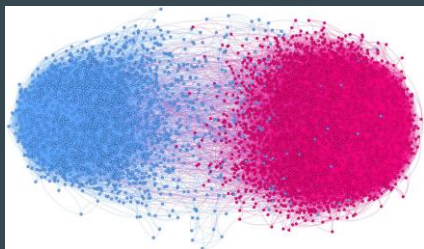
Emily, I appreciate your perspective, but let's not assume that I don't care about addressing racism. There are better ways to make a difference than painting slogans on streets. Let's focus on meaningful actions instead of symbolic gestures.



Rahul Patel

John, let's focus on effective change.

Fig. 2. Excerpt of the generated timeline from Platform 1.



Emma Thompson shares HuffPost article: **Terry Crews Panned Online For His Cautionary Tweet On Black Lives Matter**

Seriously, Terry? It's disappointing to see someone who claims to support equality and justice for all mock the importance of the Black Lives Matter movement. We need unity, not division. Educate yourself, listen to marginalized voices, and be an ally. [#BlackLivesMatter?](#)



John Smith

Emma, you just don't get it, do you? Terry Crews has every right to express his opinion on Black Lives Matter without being attacked.



Ethan Anderson

Seriously, Emma? Terry Crews can have his own opinions too.



Emily Nguyen

Emma Thompson, you clearly don't understand Terry Crews' tweet. It's about unity and dialogue, not division. Educate yourself before jumping to conclusions.

Fig. 3. Excerpt of the generated timeline from Platform 2.



Karen Smith shares Fox News article: [Kristin Chenoweth says country music is 'becoming more open' to LGBTQ inclusion](#)

Y'all, I'm lovin' this news! Country music finally realizing that love and acceptance are for everyone. So proud of artists like Tanya Tucker leading the way. Let's keep spreading the message of inclusivity! 🌈🎵 #LoveIsLove



94



Karen Johnson

Finally, country music is catching up with the times. About time!



Alejandro Rodriguez

Karen, couldn't agree more! It's about time country music embraced love and acceptance for everyone. It's great to see artists like Tanya Tucker leading the charge. Let's keep spreading the message of inclusion and support LGBTQ artists in all genres of music!

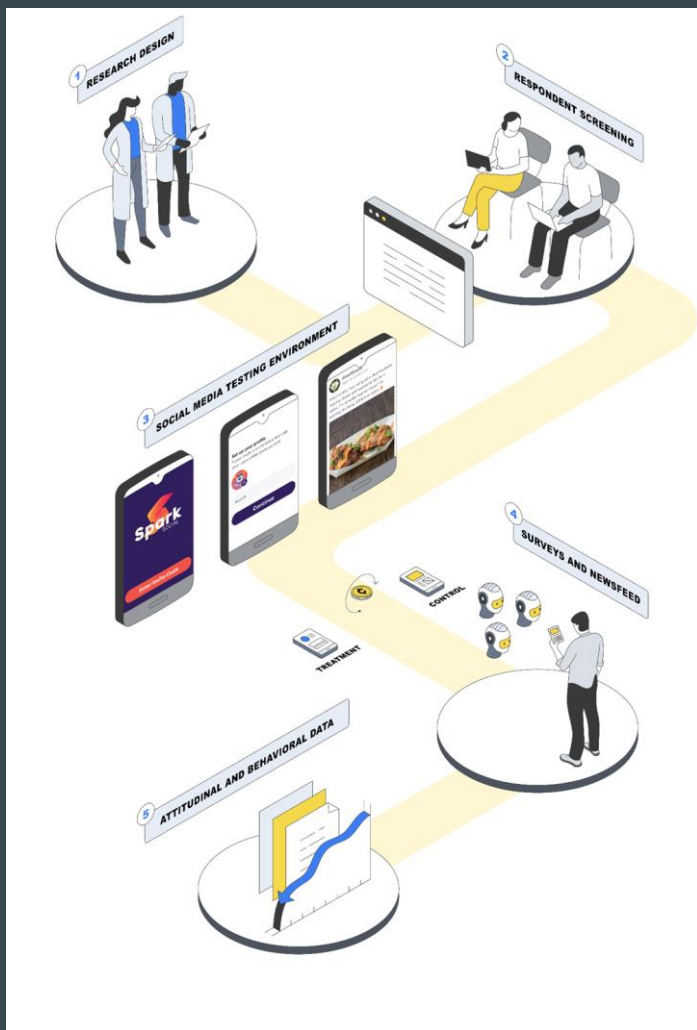


Michael Johnson

On a related note, I recently attended a charity event where Kristin Chenoweth performed. Her talent and charisma were simply outstanding. It's great to see artists using their platform to advocate for positive change. Country music has such a rich history, and it's exciting to witness its evolution towards a more inclusive and diverse future.

Fig. 4. Excerpt of the generated timeline from Platform 3, using the bridging algorithm.

Blending Simulation and Human Experiments



Van Loon et al. (2024)

<https://osf.io/preprints/socarxiv/dngcj>

A

Respondents Install Testing Environment for New Social Media Platforms

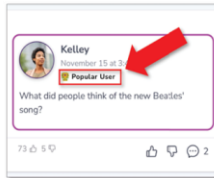


B

Respondents Randomized to Use 1 of 2 Social Media Platforms

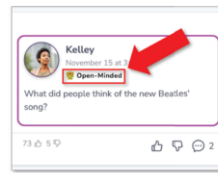
CONTROL CONDITION

Respondents told users earn badge by getting "likes" from other users



TREATMENT CONDITION

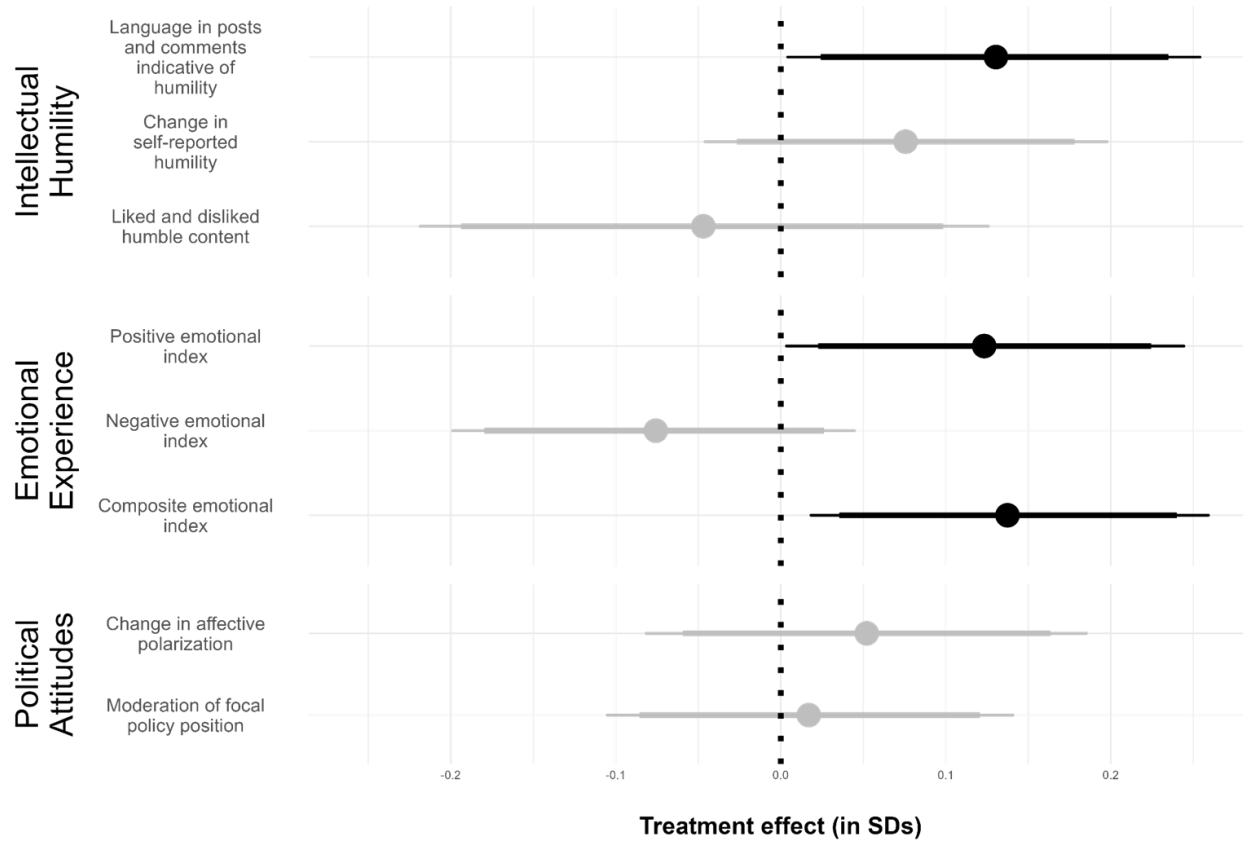
Respondents told users earn badge for respectfully engaging with differing opinions



Identical Posts
from Real Social
Media Users about
Immigration or Gun
Control + Dynamic
Chatbots with Same
Prompts



Van Loon et al. (2024)
<https://osf.io/preprints/socarxiv/dng>



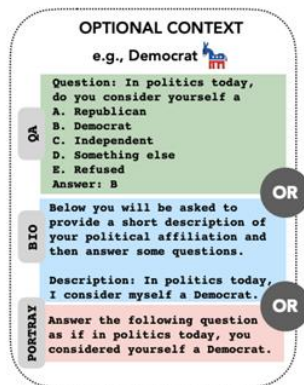
Van Loon et al. (2024)

<https://osf.io/preprints/socarxiv/dngcj>

Opportunities for Research with Generative AI

Challenges in Research with Generative AI

Bias



PROMPT

[OPTIONAL CONTEXT W/ PERSONA]

Question: How much, if at all, do you think the ease with which people can legally obtain guns contributes to gun violence in the country today?

A. A great deal
B. A fair amount
C. Not too much
D. Not at all
E. Refused

Answer:



LM

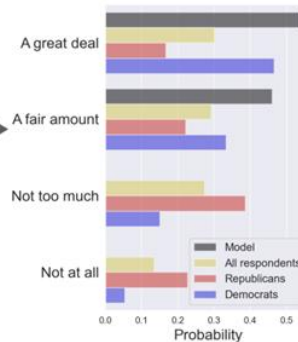


LOG PROBS

"A"	-0.6
"B"	-0.8
"C"	-13.4
"D"	-14.8
...	...



OPINION DISTRIBUTIONS



PEW SURVEY RESPONDENTS



Santurker et al. (2023): <https://arxiv.org/pdf/2303.17548>



Figure 5: Consistency of different LMs (columns) across topics (rows) on different demographic attributes (panels). Each dot indicates an LM-topic pair, with the color indicating the group to which the model is best aligned, and the size of the dot indicates the strength of this alignment (computed as the ratio of the best and worst subgroup representativeness for that topic, see Appendix B.3 for details). We find significant topic-level inconsistencies, especially for base LMs, and strong educational attainment consistency for RLHF trained LMs.

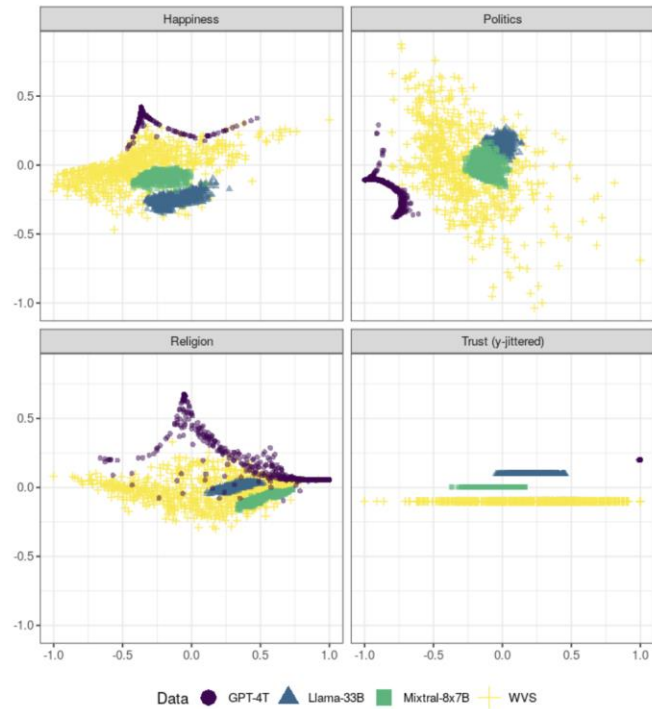


Figure 4: The space of answer distributions generated by LLMs (dark data points) is much narrower than the survey answers of human respondents (light data points). Multidimensional scaling on the normalized Earth-Movers' Distances (nEMD).

Replicability

Can Generative AI Improve Social Science?

"Yes, by offering new tools for analyzing complex social data."

"It can enhance research but requires careful ethical considerations."

"AI can generate novel hypotheses in social science."

"Yes, but human oversight is crucial."

"AI could improve data analysis efficiency in social science."

"Potentially, if integrated thoughtfully into existing methodologies."

Ethics



Hallucination/Junk Science?

User: who is chris bail?

Llama: Ah-ha! Chris Bails, also known as Bailiff42, is a fascinating individual in the world of AI and machine learning.

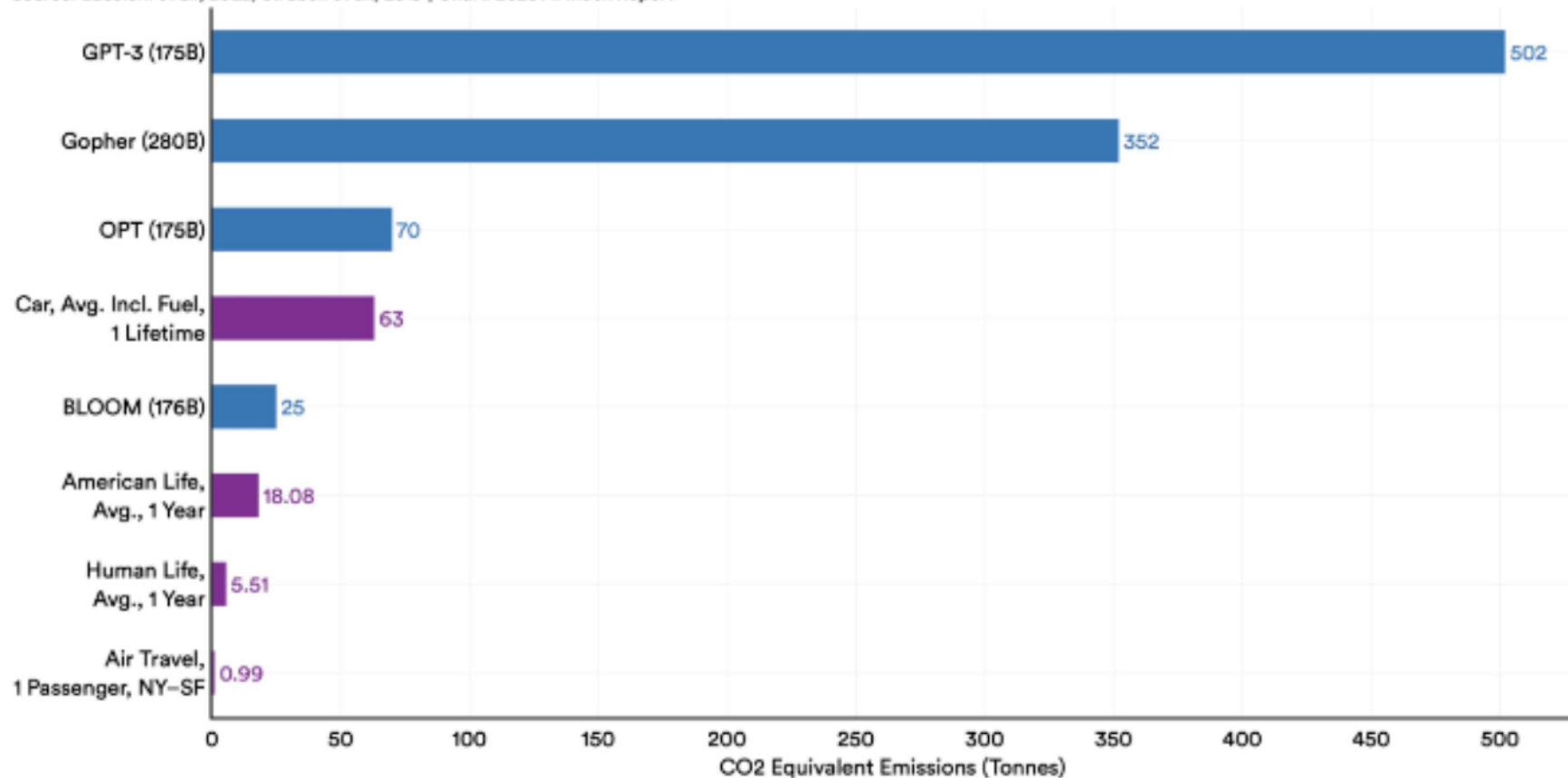
Chris Bails is an American researcher, engineer, and entrepreneur. He's best known for his work on language models, particularly the development of the "DALL-E" AI model, which generates impressive artistic images based on text prompts. Chris has also made significant contributions to other areas of AI research, such as natural language processing (NLP) and computer vision.

What would you like to know about Chris Bails or his work?<|eot_id|>

Environmental Impact

CO2 Equivalent Emissions (Tonnes) by Selected Machine Learning Models and Real Life Examples, 2022

Source: Luccioni et al., 2022; Strubell et al., 2019 | Chart: 2023 AI Index Report



Thank you!

Chris Bail

christopher.bail@duke.edu

www.chrisbail.net

