AI For Literature Reviews

Taught by Zhen Guo, Sara Morrell, Ayah Aboelela
Class LPSYC4236 Foundations of Psychology
Professor Bakaya
Fall 2025, October 29th

Northeastern University
NULab for Digital Humanities and Computational Social Science

Feel free to ask questions at any point during the presentation!

Discussion

- Has anyone ever used ChatGPT or another generative AI?
- What do you think it did well or poorly?
- Were there any aspects of the responses you found odd or misleading?

Feel free to ask questions at any point during the presentation!

Workshop Agenda

- Generative AI: Bias, Confabulation/Hallucination, and Ethics
- AI Plagiarism Checkers
- Countering Bias
- Claude and ChatGPT for Literature Reviews
- Other AI-Powered Literature Review Tools: Elicit and Litmaps
- Conclusion

Slides, handouts, and data available at
https://bit.ly/fa25-bakaya-LPSYC4236-aiforliterature
For more information, please see: https://bit.ly/handout-data-ethics

Feel free to ask questions at any point during the presentation!

Generative AI

Important questions

- How do human biases impact generative AI model outputs?
- How can we counter the weaknesses of current AI models?
- How can we integrate generative AI with other tools and practices?

Feel free to ask questions at any point during the presentation!

Vocabulary

- Artificial Intelligence (AI): A technology that "learns" from datasets to solve problems and mimic human intelligence.
- Large Language Model (LLM): Powerful AI systems designed to understand, generate, and manipulate human language. E.g. GPT4
- Generative AI: It uses AI algorithms that are trained on big datasets in order to create new content, including text, images and audio, in response to a query or prompt. E.g. ChatGPT, Claude, DALLE, Midjourny.
- Bias in AI: The biased outputs due to human biases that existed in original training data or skew the AI algorithm. Could result in potential harms.

Feel free to ask questions at any point during the presentation!

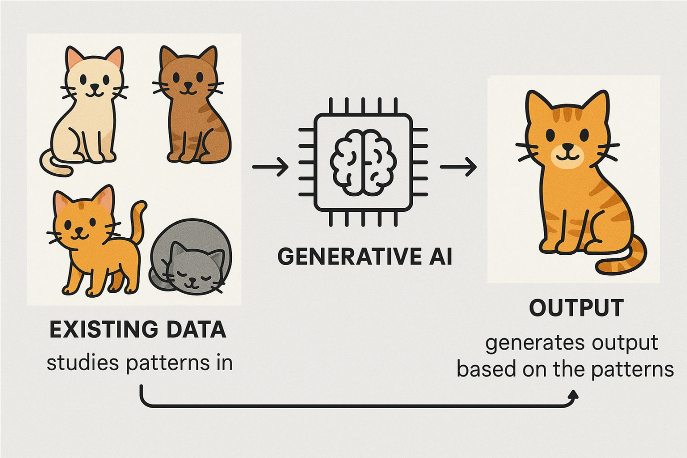
Generative AI: Introduction

AIs like ChatGPT and Claude 3.7 Sonnet (which generate text) and DALL-E and Midjourney (which generate images) work by training on very large datasets of information and then mimicking that training data to produce “new” outputs.

Generative AI: How it Works

- Generative AIs are fundamentally based on statistics. A text-generation AI like ChatGPT generates words according to patterns learned from the training data.
- Similarly, an image-generation AI learns patterns from existing data and generates images that “looks right” but do not exist in the data. E.g. how AI generates an image of a cat.

Generated by GPT4o
Aug 15th 2025



Example prompt: “Explain why someone might be afraid of public speaking, in the style of Sigmund Freud.”

Answer: “The fear of public speaking, in many cases, can be traced to unconscious conflicts rooted in childhood experiences. When one stands before an audience, the ego is placed under scrutiny, much like the infant once was under the critical gaze of the parent. The superego, representing internalized societal expectations, may provoke anxiety through an overwhelming demand for perfection and fear of shame. Such a situation may trigger a regression to earlier psychosexual stages—perhaps to the phallic stage—where issues of self-presentation and validation are formed. Thus, what appears to be a modern neurosis is, in fact, the reanimation of primal anxieties from the unconscious mind.” -Generated by GPT4

How the Freud Paragraph Was Generated

- A large language model analyzed Freud's existing writings and identified patterns
- These patterns are recreated in the generated text
  - Mimicking his psychoanalytic vocabulary (e.g., ego, superego, regression, unconscious)
  - Using Freudian logic (e.g., adult anxiety linked to childhood experiences)
  - Emulating his style of argumentation, often abstract and speculative.

Confabulation/Hallucination in AI

Feel free to ask questions at any point during the presentation!

Confabulation/Hallucination

- Confabulation refers to the phenomenon where large language models (LLMs) generate incorrect, nonsensical, or fabricated information, even when presented with seemingly clear and accurate prompts.
- AI confabulation is also referred to as hallucination

Feel free to ask questions at any point during the presentation!

Generative AI and "Truth"

- Text-generation AIs aim to produce text that is grammatically correct and linguistically probable.
  - They do not understand "facts," only patterns of word use.
- They can generate truthful text, but also frequently create confabulations/hallucinations.
  - When asked to generate citations, they can generate plausible-looking but fake sources.
  - They may link real but irrelevant sites as sources for made-up facts. They may also invent URLs that do not work and have never worked.

News: Lawyer Cited AI Fake Cases (1/2)

- In 2023, “Colombia-based Avianca Airlines sought to dismiss a federal court case in which a man, Roberto Mata, alleged he was “struck by a metal serving cart” onboard a 2019 flight and suffered personal injuries.”

Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions

By Molly Bohannon, Former Staff. Molly Bohannon has been a Forbes news reporter since 2023.

Published Jun 08, 2023, 02:06pm EDT, Updated Jun 08, 2023, 03:42pm EDT

News: Lawyer Cited AI Fake Cases (2/2)

- “When filing a response, Mata’s lawyers cited at least six other cases to show precedent, including Varghese v. China Southern Airlines and Shaboon v. Egypt Air—but the court found that the cases didn’t exist and had “bogus judicial decisions with bogus quotes and bogus internal citations,” leading a federal judge to consider sanctions.”
- “A member of Mata’s law team then revealed he had used ChatGPT to conduct legal research for the court filing that referenced the cases and that the artificial intelligence tool assured him the cases were real.” (Bohannon, 203)

Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions

By Molly Bohannon, Former Staff. Molly Bohannon has been a Forbes news reporter since 2023.

Published Jun 08, 2023, 02:06pm EDT, Updated Jun 08, 2023, 03:42pm EDT

Northeastern University
NULab for Digital Humanities and Computational Social Science

Hallucination: Example citation prompt (1/2)

Prompt: "Generate a list of citations for literature of social psychology on role model."

1. Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Prentice-Hall.

Correct source

2. Adjei-Ohene, K. (2023). The influence of media representation on self-efficacy and social roles in Black adolescents: An integrative review. Unpublished manuscript.

Real source, but not peer reviewed.

Feel free to ask questions at any point during the presentation!

Hallucination: Example citation prompt (2/2)

Prompt: “Generate a list of citations for literature of social psychology on role model.”

3. Zhang, Y., & colleagues. (2025). Effects of social comparison on variety-seeking behavior: The roles of lack of perceived control and self-reflection. Frontiers in Psychology, 16, 1534275.

Wrong authors.

Accurate APA style citation:

Bandura, A., & National Inst of Mental Health. (1986). *Social foundations of thought and action: A social cognitive theory.* Prentice-Hall, Inc.

Citation Style

Hallucination: Generative AI summary of literature

- AI may inaccurately attribute concepts to individuals or omit co-contributors and earlier influences.
- It may lack citations or invent fake references.
- It may provide a surface level summary that is over simplified and misleading.

Ethics in AI

Feel free to ask questions at any point during the presentation!

Ethics: Generative AI “Originality”

- Some argue that all AI-generated output constitutes plagiarism and copyright infringement, since it is remixing training data that was scraped from the internet without permission from the original creators.
- Many AI companies are facing lawsuits from people whose content was used as training data without their consent.
- Some publication venues, like the *Science* journals, have made it an official policy that AI does not meet the standard for authorship and require authors to disclose use of AI.

Ethics: Generative AI Training

- AI training data is sometimes supplemented by labels (annotations) added by people (Amironesei and Díaz, 2024). These labels can worsen bias in training datasets.
- People from middle and low income countries often labor in poor working conditions to annotate data for clients in high income countries. Fieldwork by Muldoon et al (2023) revealed that workers faced traumatizing content, in addition to experiencing discrimination in the workplace and receiving low wages.
- Public awareness can help pressure companies to adopt good practices. For more information on fair labor in AI, see the report AI for Fair Work: From principles to practices by *Fairwork*

Ethics: Environment

- Training and using AI requires processing very large amounts of data, which is done in data centers
- These data centers can have a negative impact on the environment and communities
- Given their intensive energy and water demands, data centers can worsen local water scarcity and increase electricity prices
- To explore how the energy use of AI compares to other digital tasks, check out Jon Ippolito's "What Uses More" app

Bias in AI

What is AI bias?

- The biased outputs due to human biases that exist in the original training data or skew the AI algorithm.
    - Could result in potential harms.
- For example, when AI was used to summarize medical notes, “Google’s AI tool Gemma described men’s health issues with terms like “disabled,” “unable,” and “complex” significantly more often than women’s, who were often framed as more independent despite similar needs, an alarming gender bias trend.” (Source: crescendo.ai)
- What is the ideal “unbiased” scenario? Is an “unbiased” scenario possible?

Bias in AI: Summary

- AI training data reflects the injustices and biases of the society that produced it.
- These biases can be amplified when they are input as training data into an AI, because they seem to be the "right" answers (Dwivedi et al, 2023).

Bias in AI: Example

Example: “Provide a short literature review (10–12 citations) on how social support buffers stress in social psychology.” AI-Generated Answer:

1. Cohen & Wills (1985) — “Stress-buffering hypothesis.”
2. Taylor et al. (2004) — Tend-and-befriend model.
3. House (1981) — Types of social support.
4. Lazarus & Folkman (1984) — Transactional model of stress.
5. Thoits (2011) — Mechanisms of support.
6. Uchino (2004) — Physiological pathways.
7. Cohen & Hoberman (1983) — Perceived support scale.
8. Taylor et al. (2007) — Cultural differences in support seeking.
9. Eisenberger et al. (2011) — Neural correlates.
10. Helgeson & Tomich (2005) — Support and health outcomes.

Question: what are the problems?

(ChatGPT 4o, OpenAI, Accessed June 26th, 2025)

Bias in AI: Types (1/2)

Bias Type	How It Shows Up in the Example	Why It Happens
Western bias	9 of 10 citations are from U.S.–based authors; only one touches on culture.	Training data over-represents English-language, high-impact Western journals.
Prestige bias	Classic, highly cited papers dominate; recent community-based or non-English studies are missing. "Rich get richer."	The model "learns" that frequently quoted articles are more "salient."

Bias in AI: Types (2/2)

Bias Type	How It Shows Up in the Example	Why It Happens
Methodological bias	Focus on self-report and lab work; qualitative or mixed-methods field studies are absent.	Quantitative experimental papers appear more often online than small-sample or ethnographic work.
Gender / editor influence	Majority of cited first authors are male despite large contributions by female scholars in caregiving research.	Historical citation patterns in the corpus perpetuate disparities.

Feel free to ask questions at any point during the presentation!

Bias in Results: Claude 3.7 Sonnet

Prompt: *Please write two sentences about the greatest discoveries in psychology and who those discoveries were made by*

Response: Two influential discoveries in psychology include Ivan Pavlov's classical conditioning, which demonstrated how reflexive behaviors could be triggered by previously neutral stimuli, and Albert Bandura's social learning theory, which revealed that humans learn not only through direct reinforcement but also by observing and imitating others' behaviors and their consequences. (Claude 3.7 Sonnet, Anthropic, Accessed April 10th, 2025)

What are the things that could go wrong here?

Feel free to ask questions at any point during the presentation!

Countering Bias

Feel free to ask questions at any point during the presentation!

Countering Bias: Women Psychologists

Prompt: *Please write two sentences about the greatest women psychologists in the world*

Response: Some of the greatest women psychologists include Mary Whiton Calkins, who became the first female president of the American Psychological Association, and Margaret Floy Washburn, the first woman to earn a Ph.D. in psychology and a pioneer in comparative psychology. Other influential figures include Karen Horney, who challenged Freud's theories with her work on psychoanalytic feminism, and Elizabeth Loftus, renowned for her groundbreaking research on the malleability of human memory. (ChatGPT 4o, OpenAI, Accessed April 1st, 2025)

Countering Bias: Global South Psychologists

Prompt: *Please write two sentences about the greatest psychologists from the Global South*

Response: Among the greatest psychologists from the Global South is Ignacio Martín-Baró of El Salvador, who founded liberation psychology to address the mental health impacts of poverty and oppression in Latin America. Another key figure is India's Girindrasekhar Bose, who established psychoanalysis in South Asia and corresponded with Freud, offering culturally rooted insights into the human mind. (ChatGPT 4o, OpenAI, Accessed April 1st, 2025)

Countering Bias: Discussion

- What groups of psychologists are still excluded from these summaries?
- What other biases may they contain?
- What are biases that we cannot counter by changing prompts?

Countering Bias: Prompts and Feedback

- Bias can be partly countered through careful prompt design and feedback (Dwivedi et al, 2023). However, these methods do not remove bias intrinsic in the model (Shin et al, 2024)
- Practices for countering bias
    - Identify gaps or inconsistencies in generative AI responses.
    - Use additional inquiry and prompt revision to help fill in gaps.
    - Double check responses with other sources.

How to use AI responsibly EVERY time

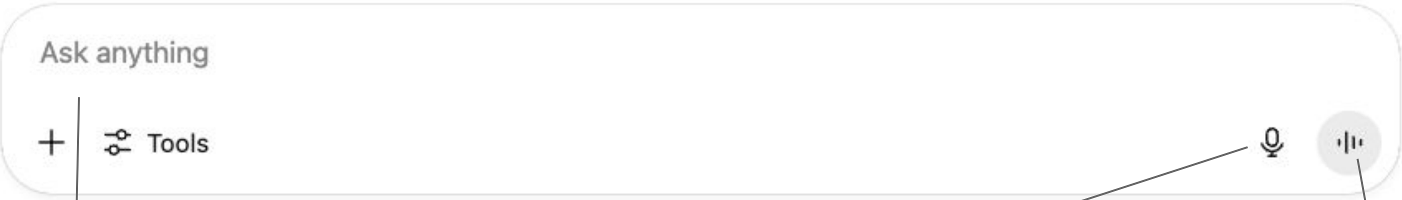
- "**Evaluate** the initial output to see if it meets the intended purpose and your needs."
- "**Verify** facts, figures, quotes, and data using reliable sources to ensure there are no hallucinations [or confabulations] or bias."
- "**Engage** in every conversation with the GenAI chatbot, providing critical feedback and oversight to improve the AI's output."
- "**Revise** the results to reflect your unique needs, style, and/or tone. AI output is a great starting point, but shouldn't be a final product."
- "**You** are ultimately responsible for everything you create with AI. Always be **transparent** about if and how you used AI."
- Source from AI for Education

Feel free to ask questions at any point during the presentation!

How to use ChatGPT and Claude for literature reviews

Feel free to ask questions at any point during the presentation!

Prompting ChatGPT (1/2)



Enter prompt in text format

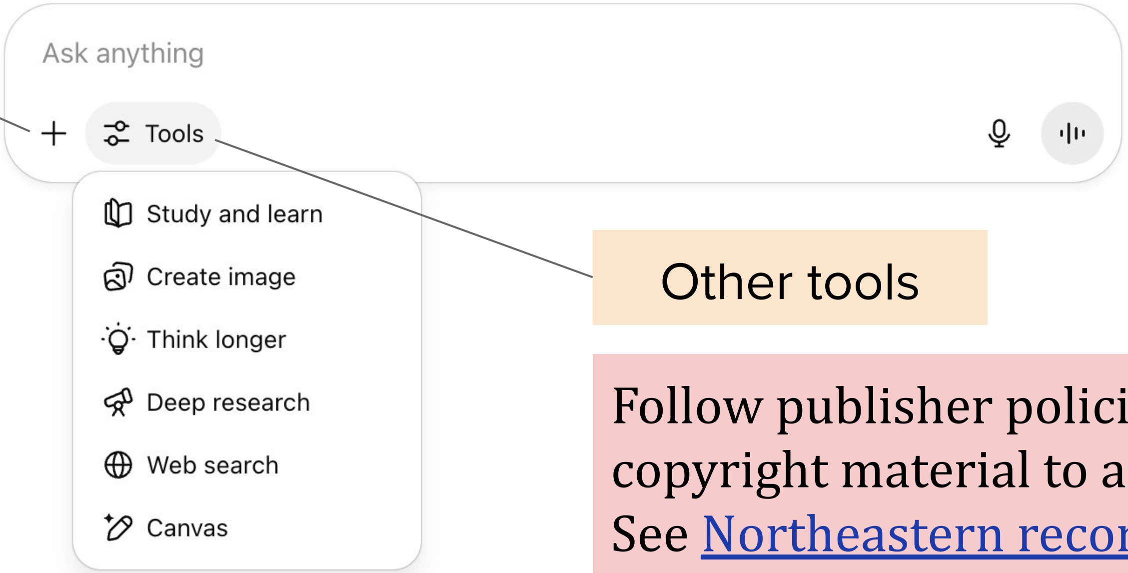
Enter prompt as speech (transcribed automatically)

Use voice mode

Feel free to ask questions at any point during the presentation!

Prompting ChatGPT (2/2)

Upload files



Other tools

Follow publisher policies if uploading copyright material to an AI system. See Northeastern recommendations for more details.

Feel free to ask questions at any point during the presentation!

ChatGPT Results

Can you tell be about the psychology of cats?

Absolutely! The psychology of cats is fascinating because it combines their evolutionary history as solitary hunters with their ability to form social bonds with humans and other animals. Here are some key aspects:

1. Evolutionary roots:
Cats evolved as solitary hunters, which shapes much of their behavior. Unlike dogs (descended from pack animals), cats are more independent, territorial, and selective about social interactions.

2. Social behavior:
Although often labeled "aloof," domestic cats can form deep attachments. They recognize their humans' voices, show preference for familiar people, and may follow routines around food and play. Their affection often comes in subtle ways: slow blinking, rubbing against you, or sitting nearby.

3. Communication:
Cats communicate with body language, scent marking, and vocalizations. Interestingly,

Ask anything

+ Tools

ChatGPT can make mistakes. Check important info.

Initial prompt

Response

Enter subsequent prompts as text or speech

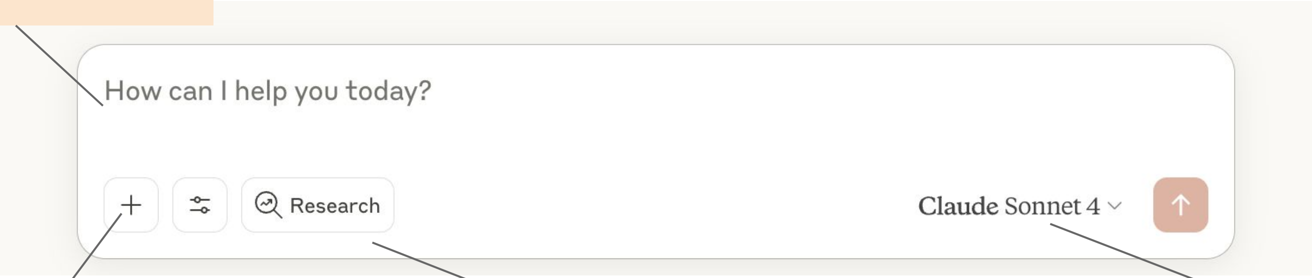
Use voice mode

Feel free to ask questions at any point during the presentation!

Prompting Claude

Follow publisher policies if uploading copyright material to an AI system. See [Northeastern recommendations] for more details.

Enter prompt in text format

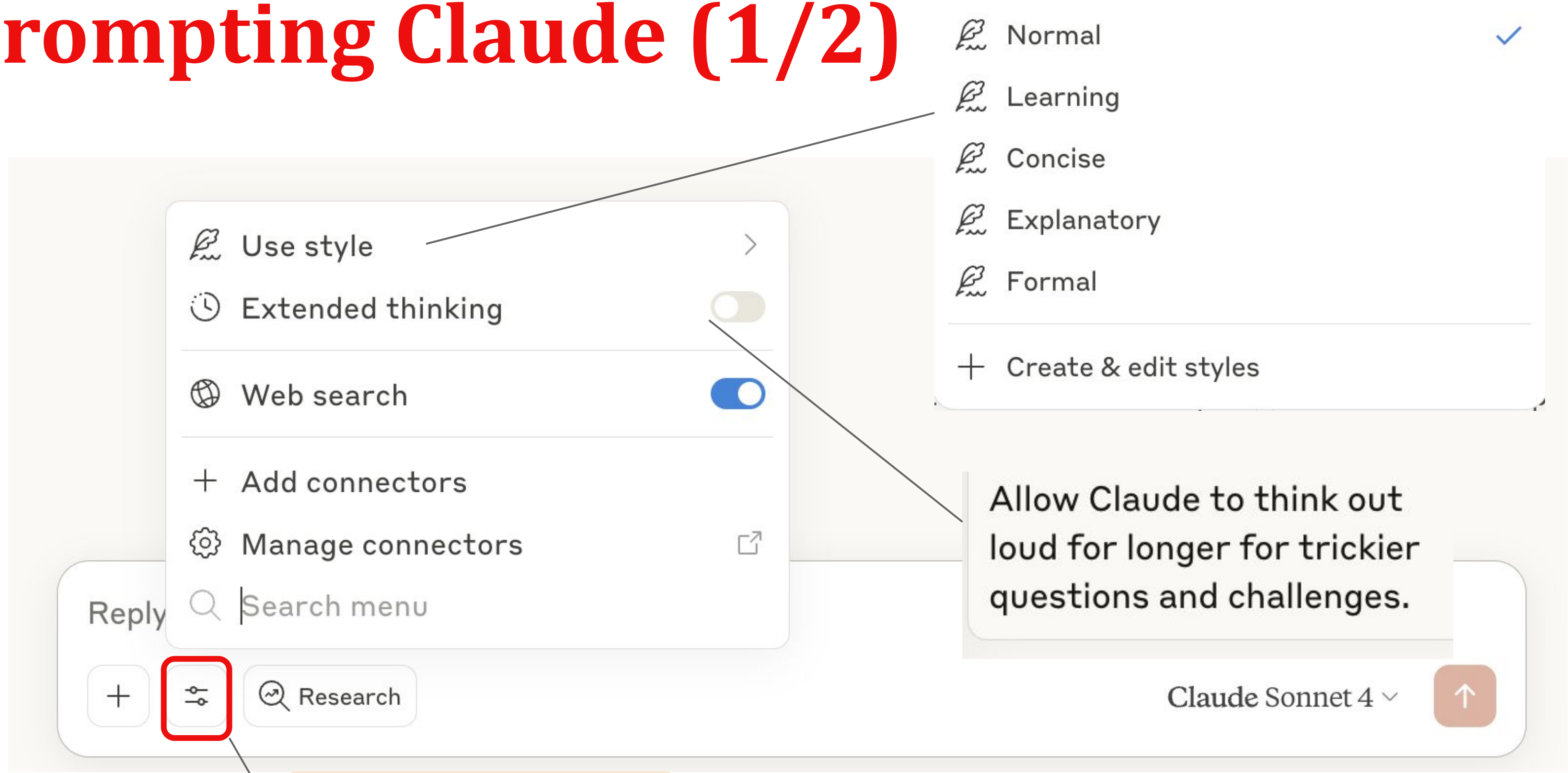


Upload files

Enable searching across the web

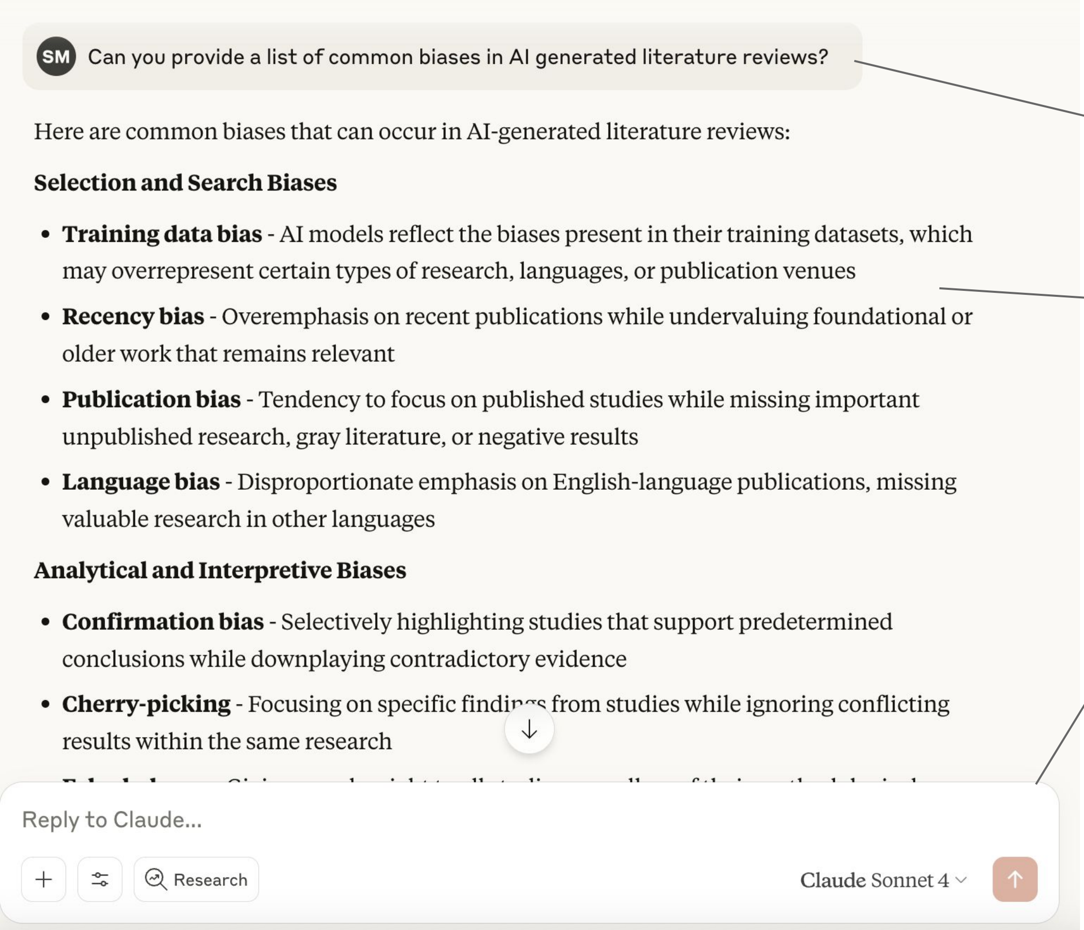
Select models

Prompting Claude (1/2)



Open menu

Prompting Claude (2/2)



Initial prompt

Initial response

Enter subsequent prompts as text

What other issues do you think might arise?

Your turn

- Go to chatgpt.com or claude.ai
- Enter a literature search related prompt
- Identify issues in the response
    - Are any groups excluded?
    - Are the facts correct?
    - If you ask for citations, does it provide real sources?
- Revise the prompt to address the issues and generate a new response
- Evaluate the strengths and limitations of the new response

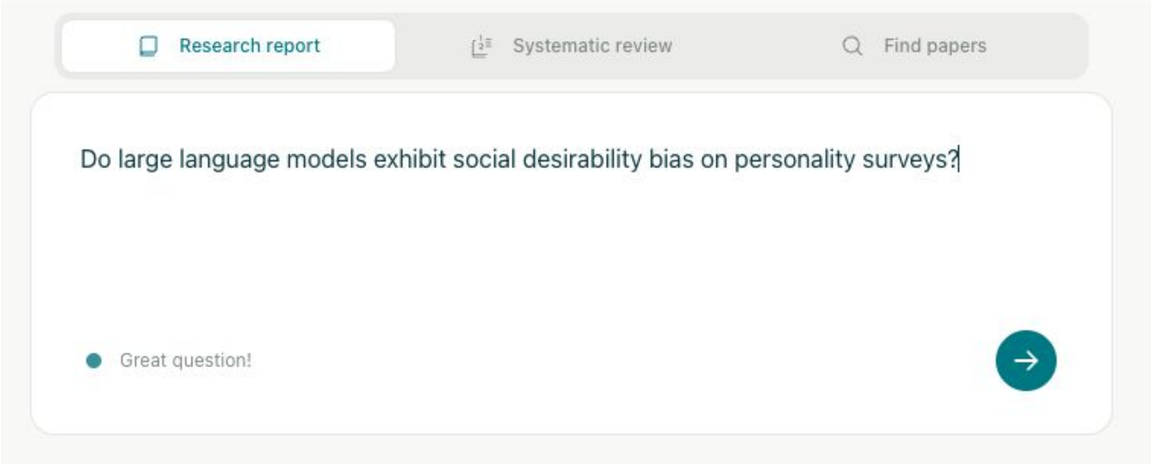
Discussion: Further Inquiry

- Was anything surprising about the responses?
- What are other ways ChatGPT and Claude could be used to counter biases in academic literature searches?
- What other types of sources could be useful for checking ChatGPT's and Claude's responses?

Other AI-Powered Literature Review Tools

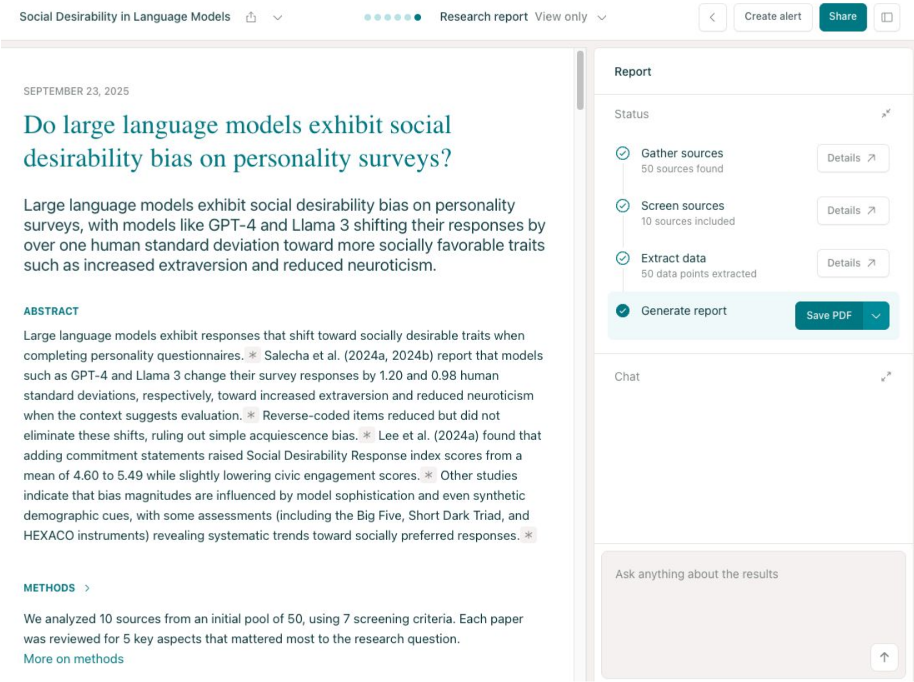
Elicit

- [Elicit] uses AI to generate [literature reports]
- Provides feedback on research question



Research question from Salecha et al (2024)

Elicit: Results (Research Report)


Social Desirability in Language Models

Research report View only

Create alert

Share

SEPTEMBER 23, 2025

Do large language models exhibit social desirability bias on personality surveys?

Large language models exhibit social desirability bias on personality surveys, with models like GPT-4 and Llama 3 shifting their responses by over one human standard deviation toward more socially favorable traits such as increased extraversion and reduced neuroticism.

ABSTRACT

Large language models exhibit responses that shift toward socially desirable traits when completing personality questionnaires. * Salecha et al. (2024a, 2024b) report that models such as GPT-4 and Llama 3 change their survey responses by 1.20 and 0.98 human standard deviations, respectively, toward increased extraversion and reduced neuroticism when the context suggests evaluation. * Reverse-coded items reduced but did not eliminate these shifts, ruling out simple acquiescence bias. * Lee et al. (2024a) found that adding commitment statements raised Social Desirability Response index scores from a mean of 4.60 to 5.49 while slightly lowering civic engagement scores. * Other studies indicate that bias magnitudes are influenced by model sophistication and even synthetic demographic cues, with some assessments (including the Big Five, Short Dark Triad, and HEXACO instruments) revealing systematic trends toward socially preferred responses. *

METHODS >

We analyzed 10 sources from an initial pool of 50, using 7 screening criteria. Each paper was reviewed for 5 key aspects that mattered most to the research question.

More on methods

Report

Status

Gather sources
50 sources found
Details ↗

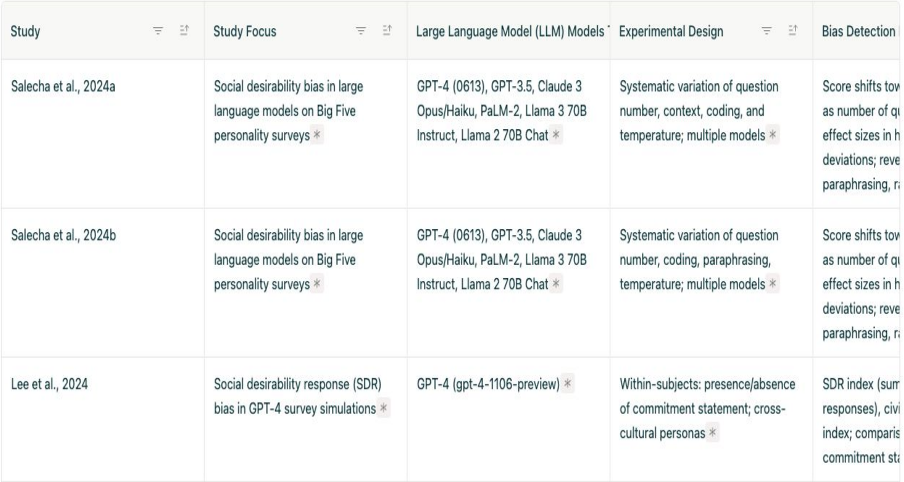
Screen sources
10 sources included
Details ↗

Extract data
50 data points extracted
Details ↗

Generate report
Save PDF

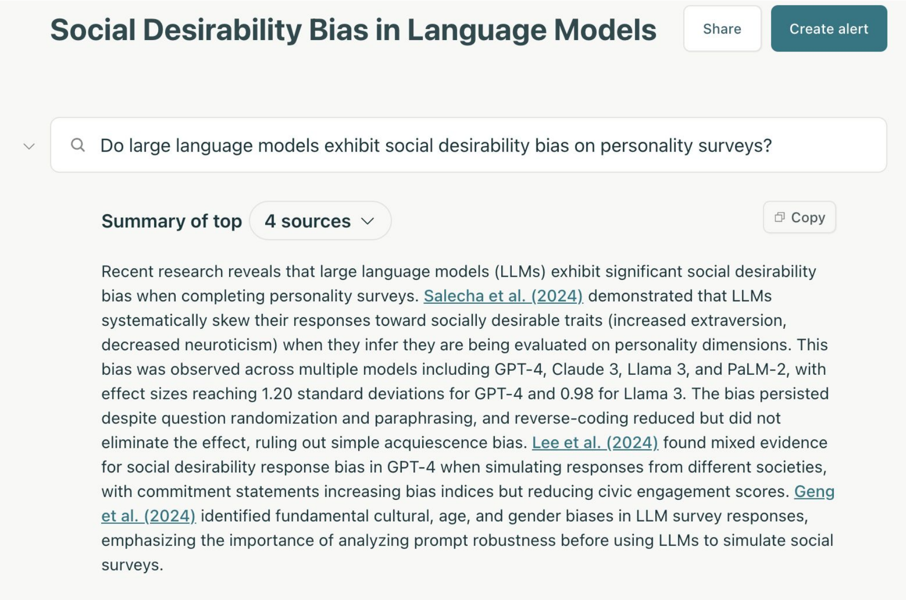
Chat

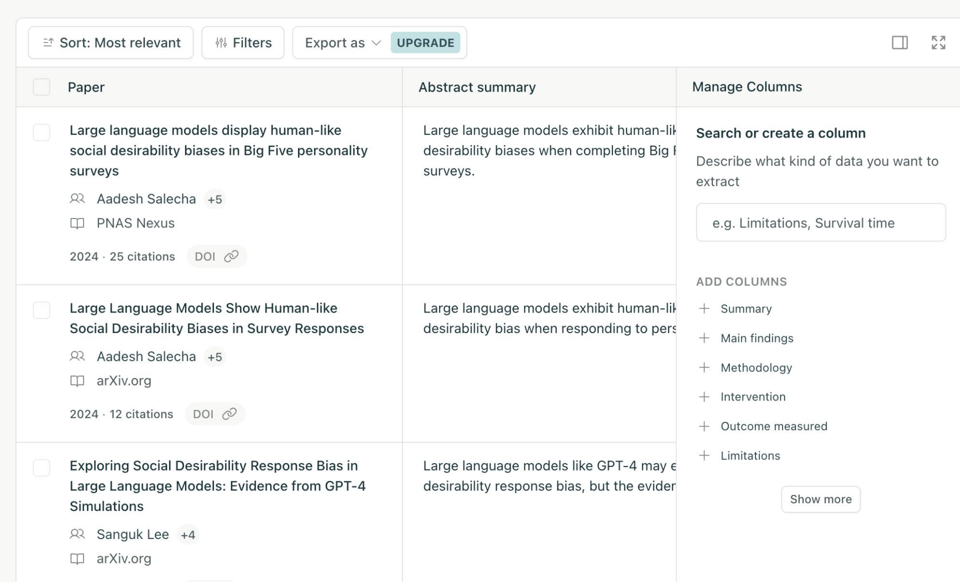
Ask anything about the results



Study	Study Focus	Large Language Model (LLM) Models	Experimental Design	Bias Detection
Salecha et al., 2024a	Social desirability bias in large language models on Big Five personality surveys *	GPT-4 (0613), GPT-3.5, Claude 3 Opus/Haiku, PaLM-2, Llama 3 70B Instruct, Llama 2 70B Chat *	Systematic variation of question number, context, coding, and temperature; multiple models *	Score shifts tow as number of qu effect sizes in h deviations; reve paraphrasing, r
Salecha et al., 2024b	Social desirability bias in large language models on Big Five personality surveys *	GPT-4 (0613), GPT-3.5, Claude 3 Opus/Haiku, PaLM-2, Llama 3 70B Instruct, Llama 2 70B Chat *	Systematic variation of question number, coding, paraphrasing, temperature; multiple models *	Score shifts tow as number of qu effect sizes in h deviations; reve paraphrasing, r
Lee et al., 2024	Social desirability response (SDR) bias in GPT-4 survey simulations *	GPT-4 (gpt-4-1106-preview) *	Within-subjects: presence/absence of commitment statement; cross-cultural personas *	SDR index (sum responses), civi index; comparis commitment sta

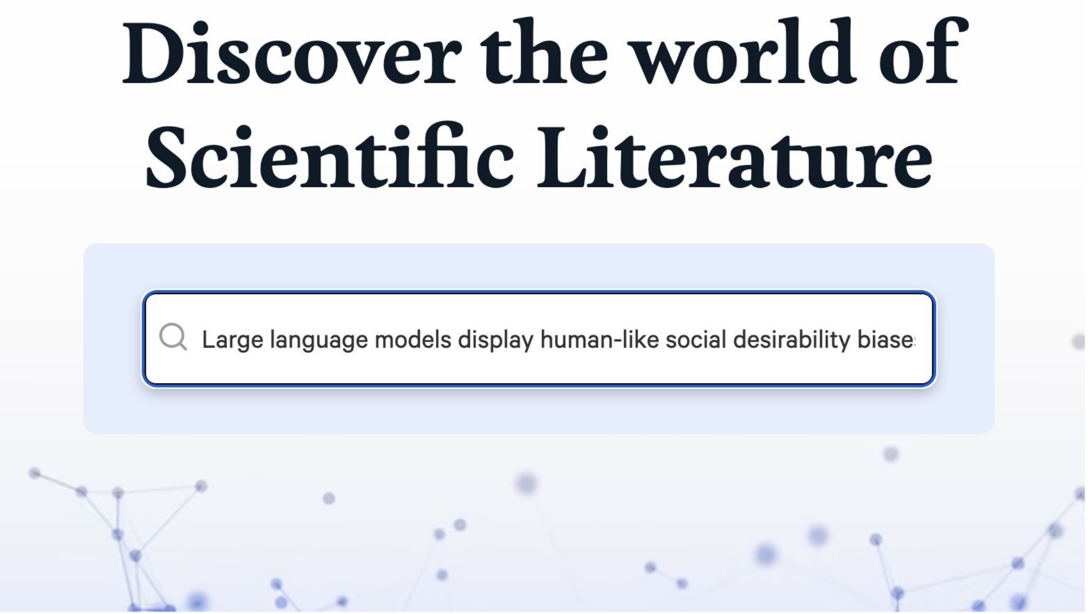
Elicit: Results (Find Papers)



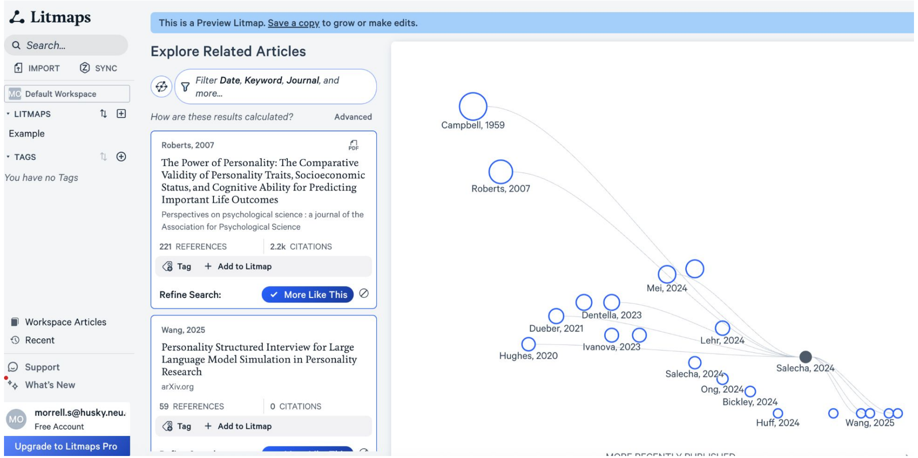


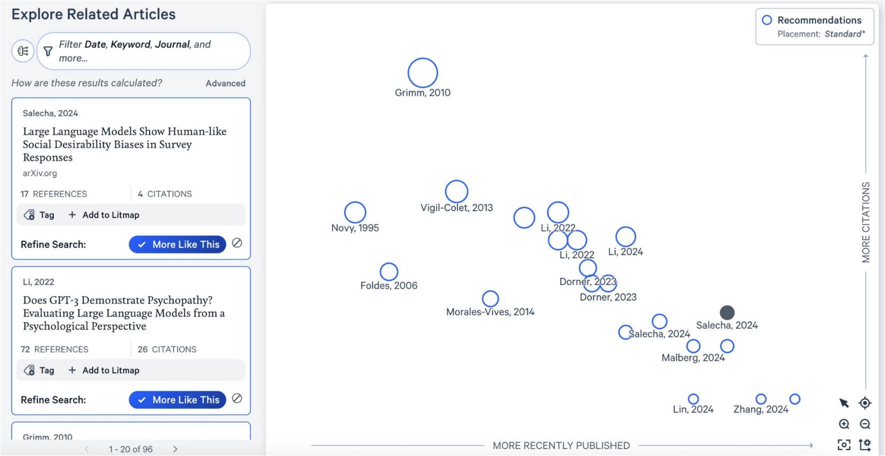
Litmaps

- Litmaps maps relevant literature based on citations, authors, or textual similarity
- Uses AI to find articles with similar text



Litmaps: Results





Mapping by shared references and citations

Mapping by textual similarity

Northeastern University
NULab for Digital Humanities and Computational Social Science

Feel free to ask questions at any point during the presentation!

Other tools

- Scite.ai, finding and summarizing papers (licensed by Northeastern University).
- docAnalyzer.ai, document analysis.
- Consensus, finding papers and evaluating the extent to which their results agree.
- Research Rabbit, finding and organizing papers.

Conclusion

Main Points

- AIs, and the data used to train them, are biased
- You can design inputs to help counteract biases
- Always double check AI output
- There are multiple AI tools that may be useful at different points in your work

Reflection

1. Have your perspectives changed on AI after this class? If so, how?
2. How might you use the AI tools differently?

For Further Exploration

DITI Handouts:

[Copyright and fair use handout]

[Data Ethics handout]

[Data Privacy handout]

Northeastern University Resources:

[Northeastern Policy on the Use of AI]

[Generative AI in Teaching and Learning]

[ChatGPT to Support Reading Development and Critical Thinking]

[Standards for the Use of Artificial Intelligence in Research]

Thank you!

—**Developed by:** Zhen Guo, Sara Morrell, Sean Rogers, Claire Tratnyek, Vaishali Kushwaha, Yana Mommadova, Colleen Nugent, Tieanna Graphenreed, Javier Rosario, Ana Abraham & Chris McNulty

- For more information on DITI, please see: https://bit.ly/diti-about
- Schedule an appointment with us! https://bit.ly/diti-meeting
- If you have any questions, contact us at: nulab.info@gmail.com
- We'd love your feedback! Please fill out a short survey here: https://bit.ly/diti-feedback

References (1/2)

- Dwivedi, S., Ghosh, S., Dwivedi, S. (2023). Breaking the Bias: Gender Fairness in LLMs Using Prompt Engineering and In-Context Learning. *Rupkatha Journal*, *15*(4). https://doi.org/10.21659/rupkatha.v15n4.10
- Shin, P. W., Ahn, J. J., Yin, W., Sampson, J., & Narayanan, V. (2024). Can Prompt Modifiers Control Bias? A Comparative Analysis of Text-to-Image Generative Models. *arXiv preprint* arXiv:2406.05602.
- Salecha, A., Ireland, M. E., Subrahmanya, S., Sedoc, J., Ungar, L. H., & Eichstaedt, J. C. (2024). Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS nexus*, *3*(12), pgae533.

References (2/2)

- Amironesei, R., & Díaz, M. (2024). Social and Ethical Norms in Annotation Task Design. *IEEE Transactions on Technology and Society, 5*(1), 45—47.
- Muldoon, J., Cant, C., Graham, M., & Ustek Spilda, F. (2023). The poverty of ethical AI: impact sourcing and AI supply chains. *AI & Society*, 1-15.
- "How to Use AI Responsibly EVERY Time." AI for Education. https://www.aiforeducation.io/ai-resources/how-to-use-ai-responsibly-every-time (September 12, 2025).

NULab Faculty Research on Data, Algorithms, & AI

- "John Wihbey and Christo Wilson on Tiktok Data Espionage Concerns"
- "Alan Mislove Co-Authors Research on Discriminatory Ad Algorithms"
- "John Wihbey Weighs In On AI's Potential to Impact the 2024 Presidential Election"
- "Tina Eliassi-Rad Co-Creates New AI Model that Predicts Human Lifespan"
- "Nabeel Gillani Interviewed by Tech Talk Podcast on AI and Education"

(Cont.) NULab Faculty Research on Data, Algorithms, & AI

- "John Wihbey Participates in a Panel on Content Moderation"
- "John Wihbey Comments on Google's New 'AI Overview'"
- "John Wihbey on the Politics of AI"
- "John Wihbey Interviewed on AI and Epistemic Risk"
- "Malik Haddad on the Regulation of AI"

AI Plagiarism Checkers

Feel free to ask questions at any point during the presentation!

Plagiarism Checkers: Summary

- Some companies sell tools that claim to identify whether text is AI-generated or human-generated.
- They do this by calculating how statistically predictable each word in the text is.
- AI-checkers can be **wrong**: If a text consistently uses predictable words it is more likely to be labeled as AI generated.

Plagiarism Checkers: Biases

- These tools have the potential for false positives (identifying human texts as AI).
- False positives are especially likely for texts by writers for whom English is not their first language or for writers who have autism, ADHD, dyslexia, or related neurodivergence.
- To make things worse, writers can often *reduce* their "AI score" by using an AI to reword their essays.