

<https://bit.ly/MSU AI Workshop 2025>

Link to slides →



AI in the Research Lifecycle

Exploring some of the possibilities, pitfalls, and politics

Zoe LeBlanc, Assistant Professor

Contact info:
zleblanc@illinois.edu / zoeleblanc.bsky.social



School of
Information Sciences
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Terms & Conditions

1. Generosity & Appreciation

- Many thanks to the organizers and all participants 
- Also, many thanks the enormous labor and generosity to all the scholars, artists, practitioners, and students working on this topic 
- Critique as care, not a takedown 

2. Definitions & Answers

- Breadth over depth 
- Very new and fast-moving landscape so hard to be comprehensive or definitive 

3. Positionality & Limitations

- Specialize in histories of information & digital humanities & work as a TT prof at an R1 so YMMV 
- Mostly presenting in this session, so hoping we can discuss more in following sessions 
- PLEASE ASK QUESTIONS & PARTICIPATE  !!

Overview of Workshop Schedule

1. Directions & Reflections

- How are academics using AI, at least so far? What are some of the trends across disciplines?
What are our research workflows and lifecycles?

2. Brainstorming & Exploring

- How might AI inform how we develop research questions and projects? How can we effectively use chatbots and how much do we need to understand AI?

3. Collecting & Curating

- How can AI help us collect and curate our research materials? How does AI fit with other existing scholarly infrastructures?

4. Designing & Analyzing

- How can we utilize AI to design our research and learn new methods? How can AI also help us analyze our results? How much can we trust AI?

5. Evaluating & Preserving

- How can we develop practices that support reproducible and sustainable research with AI?
How can we ensure that our practices are in line with our political and ethical values?

Part 1

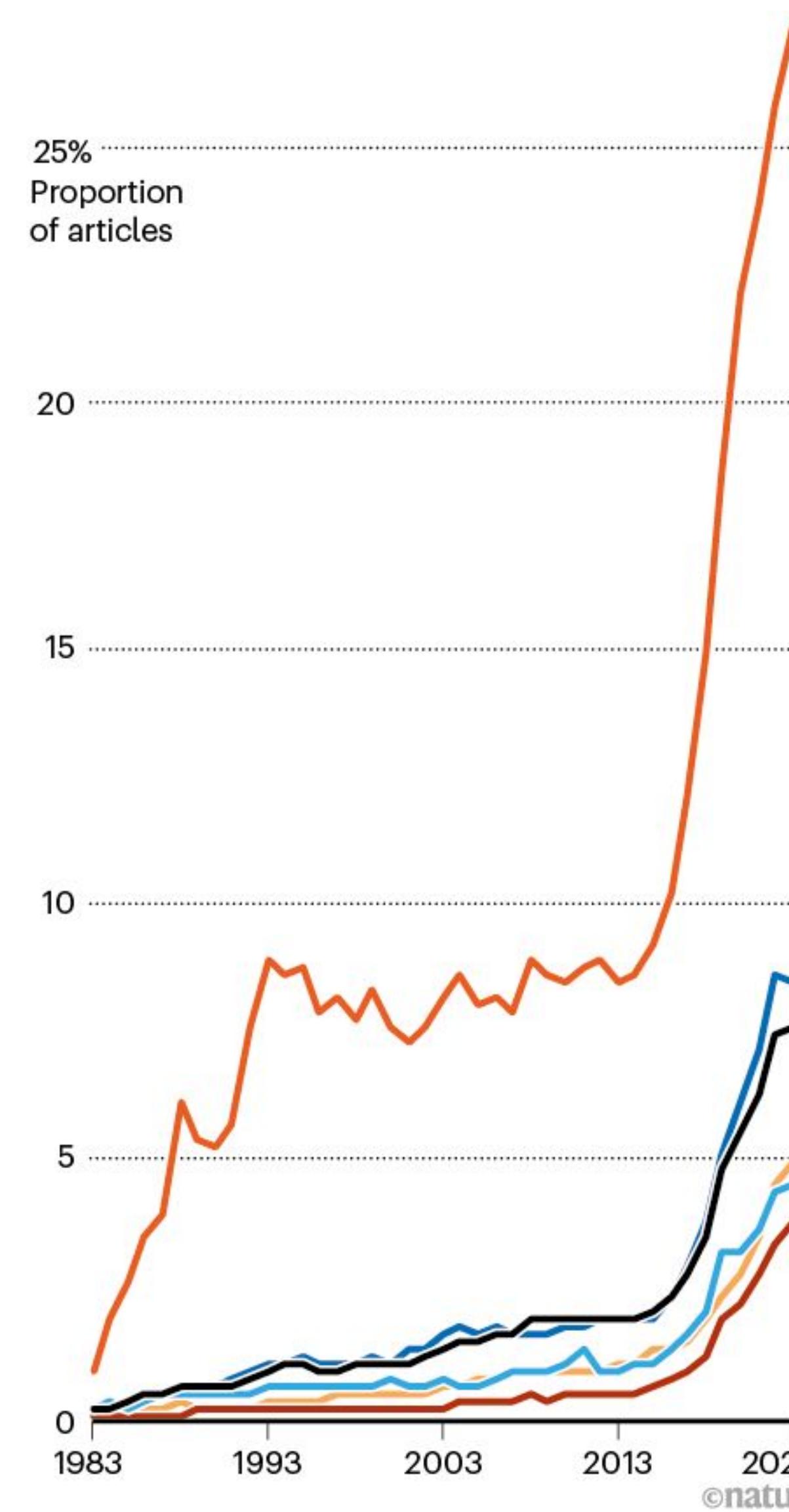
**Some Initial Directions &
Reflections To Ground Our Use
of AI in Research**

Initial Findings on AI & STEM

AI ON THE RISE

The share of research papers with titles or abstracts that mention AI or machine-learning terms has risen to around 8%, analysis of the Scopus database suggests.

- Computer science
- Physical sciences
- Life sciences
- Social sciences
- Health and medicine
- Total



USING GENERATIVE AI

Q: How often do you use generative AI tools (such as ChatGPT) at work?

- I use them every day
- I use them occasionally
- Never
- I use them more than once a week
- I've used them only a few times

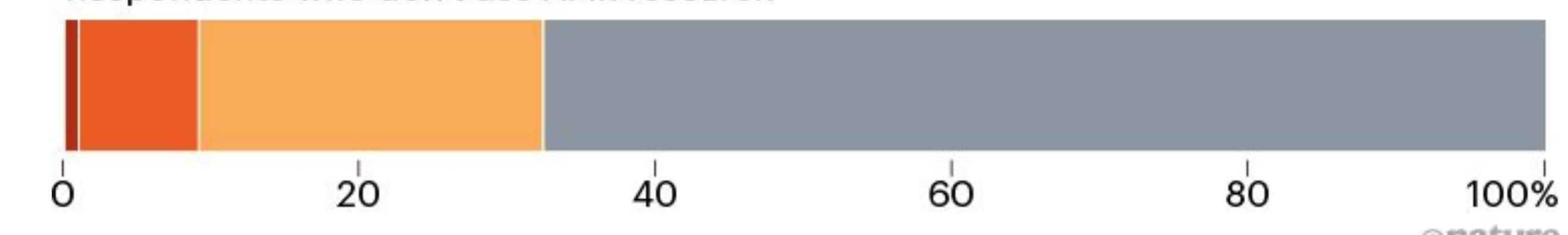
Respondents who study AI



Respondents who use AI in research



Respondents who don't use AI in research



NEWS FEATURE | 27 September 2023 | Correction 10 October 2023

AI and science: what 1,600 researchers think

A *Nature* survey finds that scientists are concerned, as well as excited, by the increasing use of artificial-intelligence tools in research.

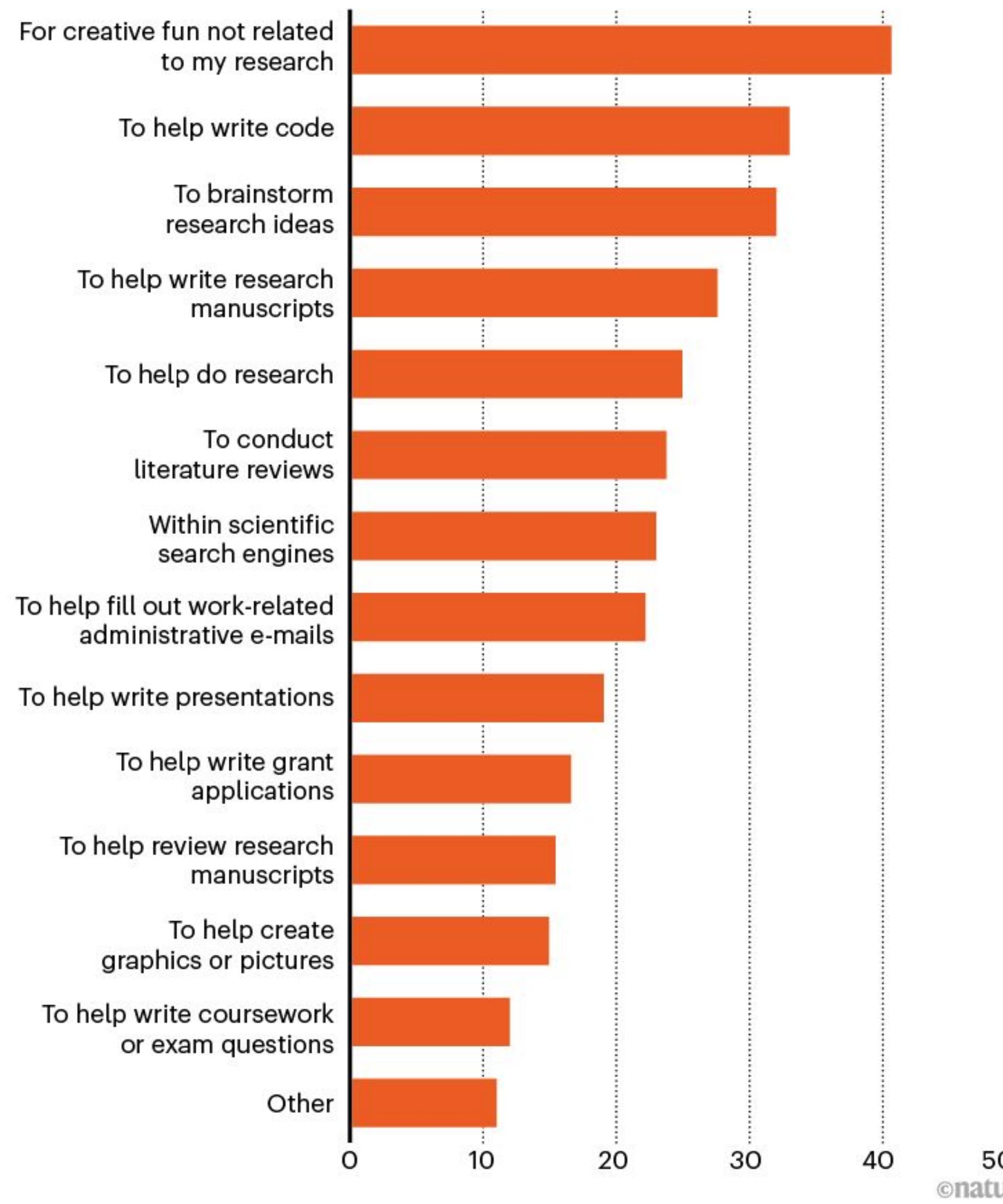
By [Richard Van Noorden](#) & [Jeffrey M. Perkel](#)

©nature

Initial Findings on AI & STEM

HOW RESEARCHERS USE LARGE LANGUAGE MODELS

Q: What do you use generative AI tools (such as ChatGPT and other large language models) for? (Choose all that apply.)



AI ANTICIPATIONS

Q: How useful do you think AI tools are for researchers in your field?

■ Essential ■ Very useful ■ Useful ■ Slightly useful ■ Not at all useful

Respondents who use AI in research



Respondents who don't use AI in research



Q: How useful do you think AI tools will become for researchers in your field in the next decade?

Respondents who use AI in research



Respondents who don't use AI in research



AI and science: what 1,600 researchers think

A *Nature* survey finds that scientists are concerned, as well as excited, by the increasing use of artificial-intelligence tools in research.

Initial Findings on AI & (Digital) Humanities

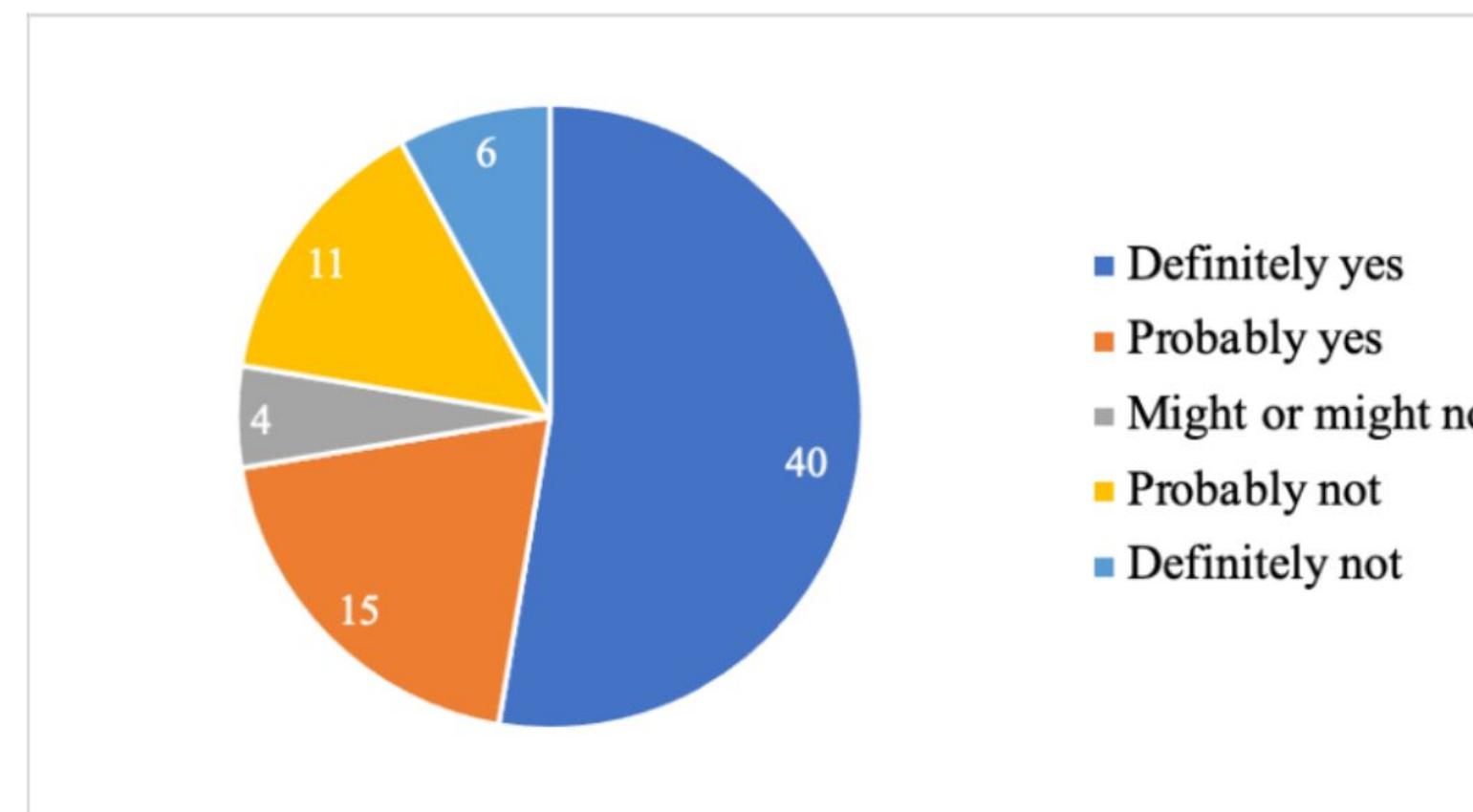


Figure 1. The frequency of self-identification as a DH scholar.

Table 1. The frequency and percentage of DH scholars' use of GenAI tools in their personal life and research.

Use of GenAI	Personal life		Research and science	
	Frequency	Percentage	Frequency	Percentage
Yes	43	73%	33	56%
Maybe	7	12%	4	7%
No	9	15%	22	37%
	59	100%	59	100%

arXiv > cs > arXiv:2404.12458

Computer Science > Artificial Intelligence

[Submitted on 18 Apr 2024]

The collective use and evaluation of generative AI tools in digital humanities research: Survey-based results

Meredith Dedema, Rongqian Ma

The advent of generative artificial intelligence (GenAI) technologies has revolutionized research, with significant implications for Digital Humanities (DH), a field inherently intertwined with technological progress. This article investigates how digital humanities scholars adopt, practice, as well as critically evaluate, GenAI technologies such as ChatGPT in the research process. Drawing on 76 responses collected from an international survey study, we explored digital humanities scholars' rationale for GenAI adoption in research, identified specific use cases and practices of using GenAI to support various DH research tasks, and analyzed scholars' collective perceptions of GenAI's benefits, risks, and impact on DH research. The survey results suggest that DH research communities hold divisive sentiments towards the value of GenAI in DH scholarship, whereas the actual usage diversifies among individuals and across research tasks. Our survey-based analysis has the potential to serve as a basis for further empirical research on the impact of GenAI on the evolution of DH scholarship.

Subjects: Artificial Intelligence (cs.AI)
Cite as: arXiv:2404.12458 [cs.AI]
(or arXiv:2404.12458v1 [cs.AI] for this version)
<https://doi.org/10.48550/arXiv.2404.12458> ⓘ

Submission history
From: Rongqian Ma [view email]
[v1] Thu, 18 Apr 2024 18:33:00 UTC (668 KB)

Initial Findings on AI & (Digital) Humanities

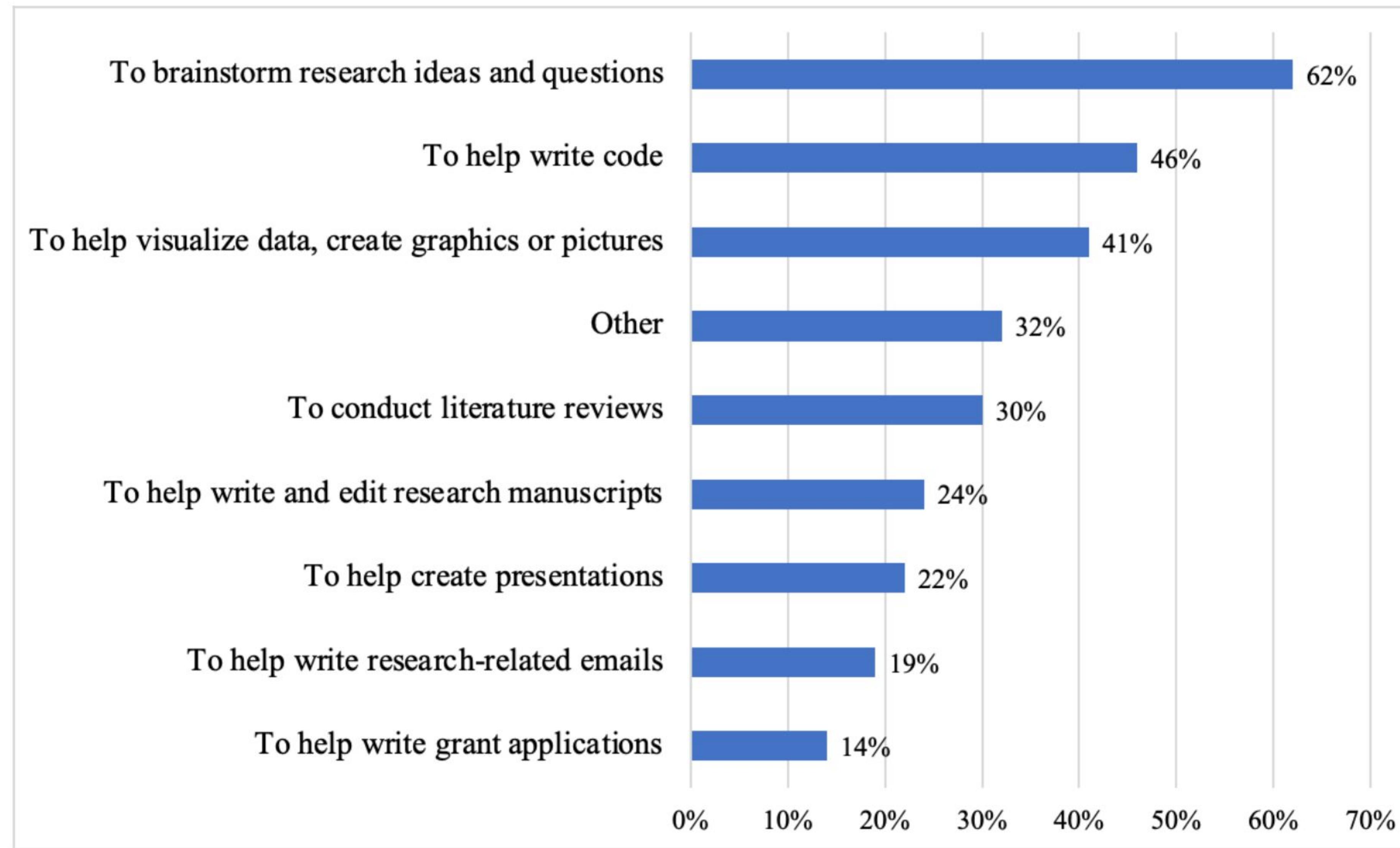


Figure 2. The purpose of using GenAI in research among DH scholars.

Initial Findings on AI & (Digital) Humanities

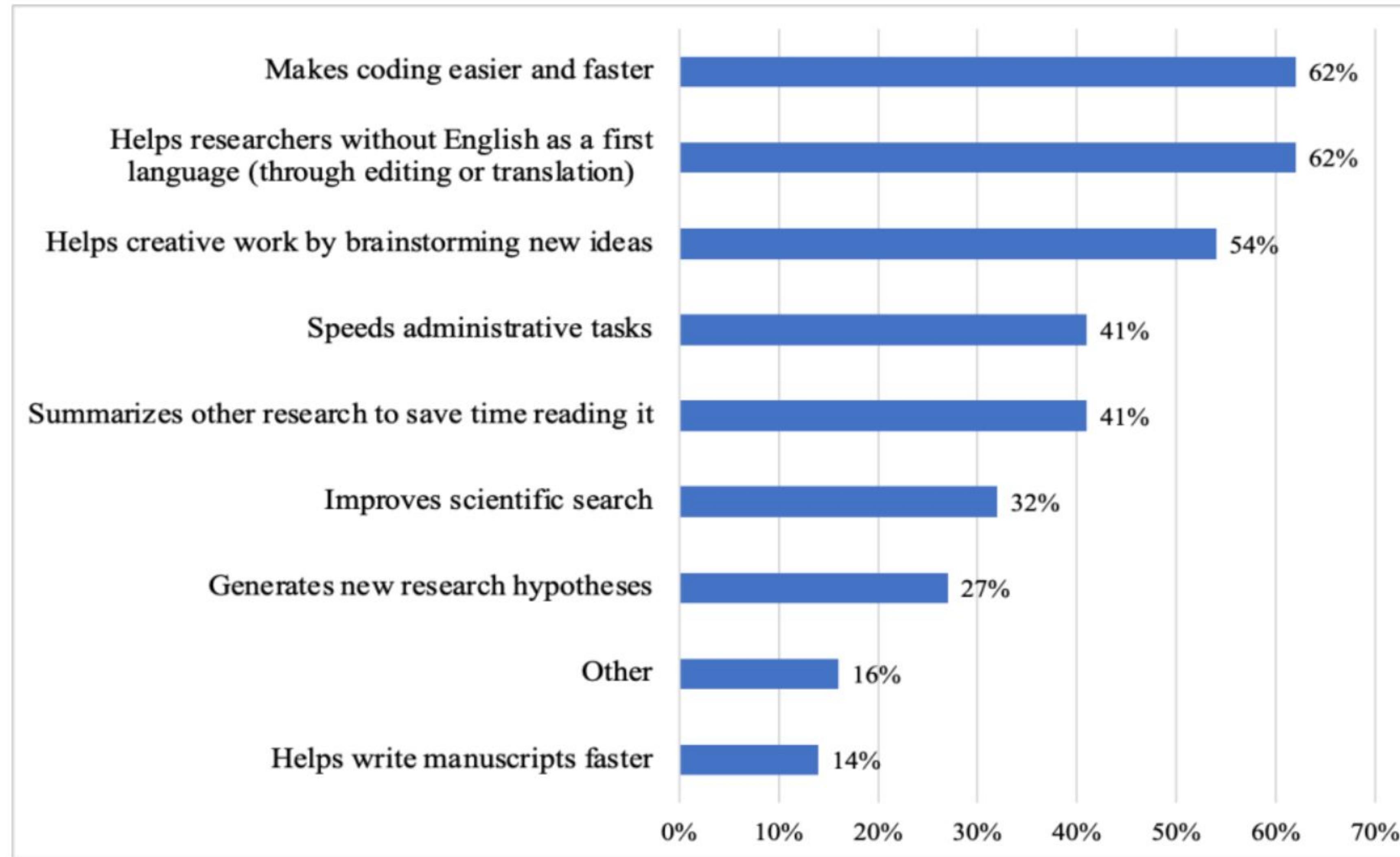


Figure 3. The benefits of using GenAI in research among DH scholars.

Initial Findings on AI & STEM/Social Sciences



[Submitted on 30 Oct 2024]

LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions

Zehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Chen, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, Amy X. Zhang

The rise of large language models (LLMs) has led many researchers to consider their usage for scientific work. Some have found benefit in using LLMs to augment or automate aspects of their research pipeline, while others have urged caution due to risks and ethical concerns. This work has sought to quantify and characterize how researchers use LLMs and why. We present the first large-scale survey of 816 verified article authors to understand how the research community leverages LLMs as research tools. We examine participants' self-reported LLM usage, finding that 81% of researchers have already incorporated LLMs into different aspects of their research workflow. We also find that traditionally disadvantaged groups in academia (non-White, junior, non-native English speaking researchers) report higher LLM usage than perceived benefits, suggesting potential for improved research efficiency. However, women, non-binary, and senior researchers have greater concerns, potentially hindering adoption.

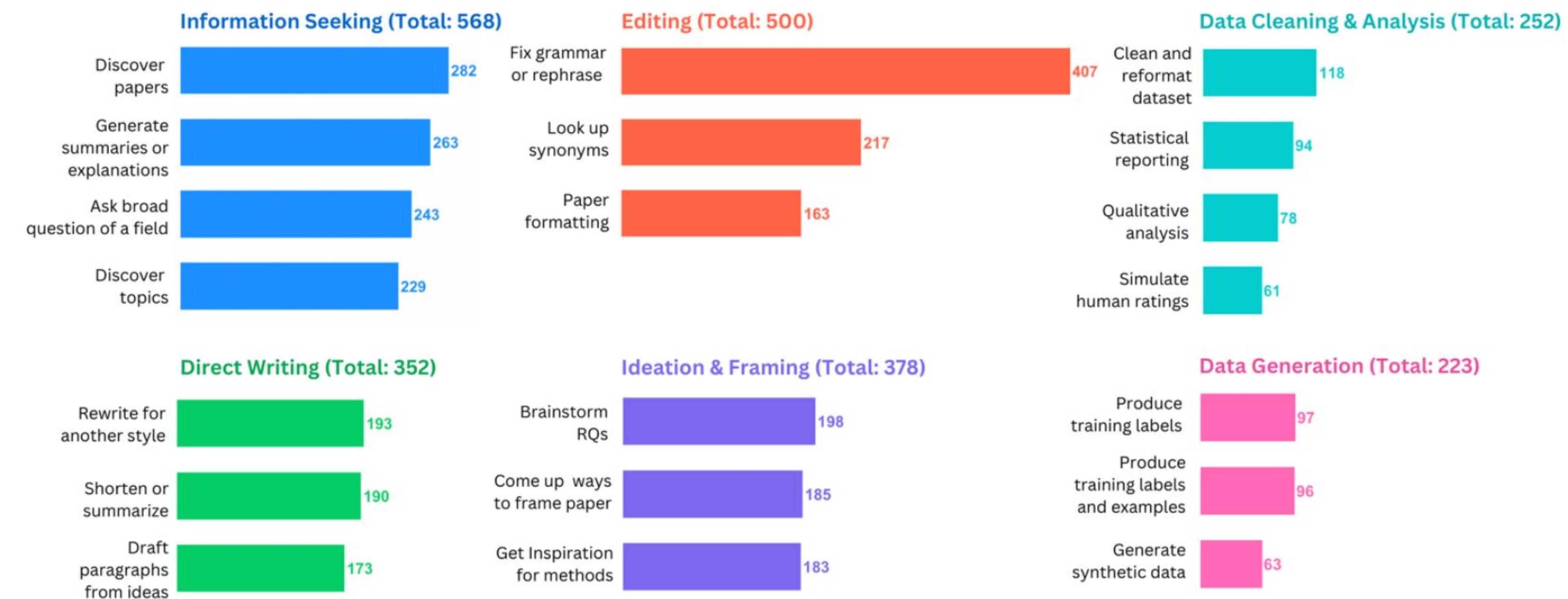


Fig. 3. **LLM Usage Breakdown Under Each Usage Type (N=816).** Each bar chart shows the number of participants who reported using LLMs for tasks that were subcategories of each usage type. Every participant could select multiple subcategories. The total number next to each title shows the number of participants who indicated using LLMs for the broad usage type.

Initial Findings on AI & STEM/Social Sciences

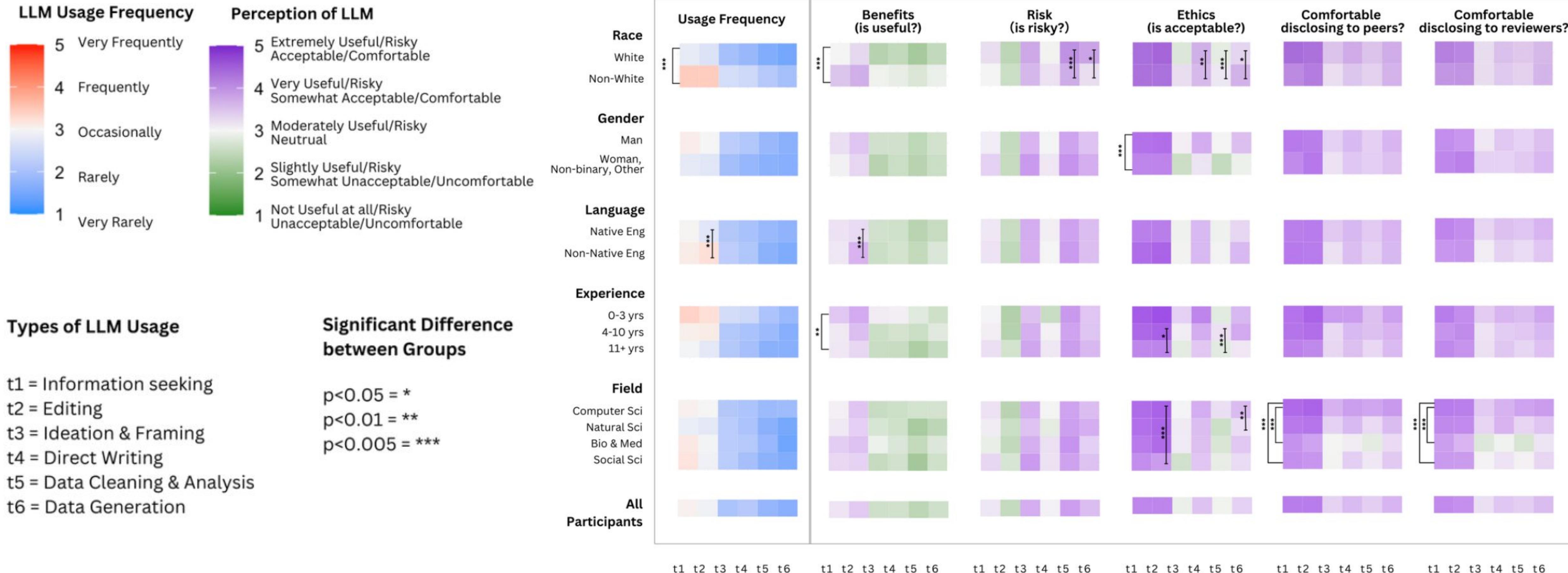


Fig. 4. Overview of our survey results (N is shown in Table 3), broken apart by demographic characteristics. Each heat map square represents the **average** rating of this demographic group on the usage frequency or perception for the particular type of LLM usage. The stars (***) are the significance levels of the differences indicated by p values from regression results. The brackets on the left indicate this difference is significant across all types of usage whereas the lines between squares indicate this difference is only significant for certain types of usage.

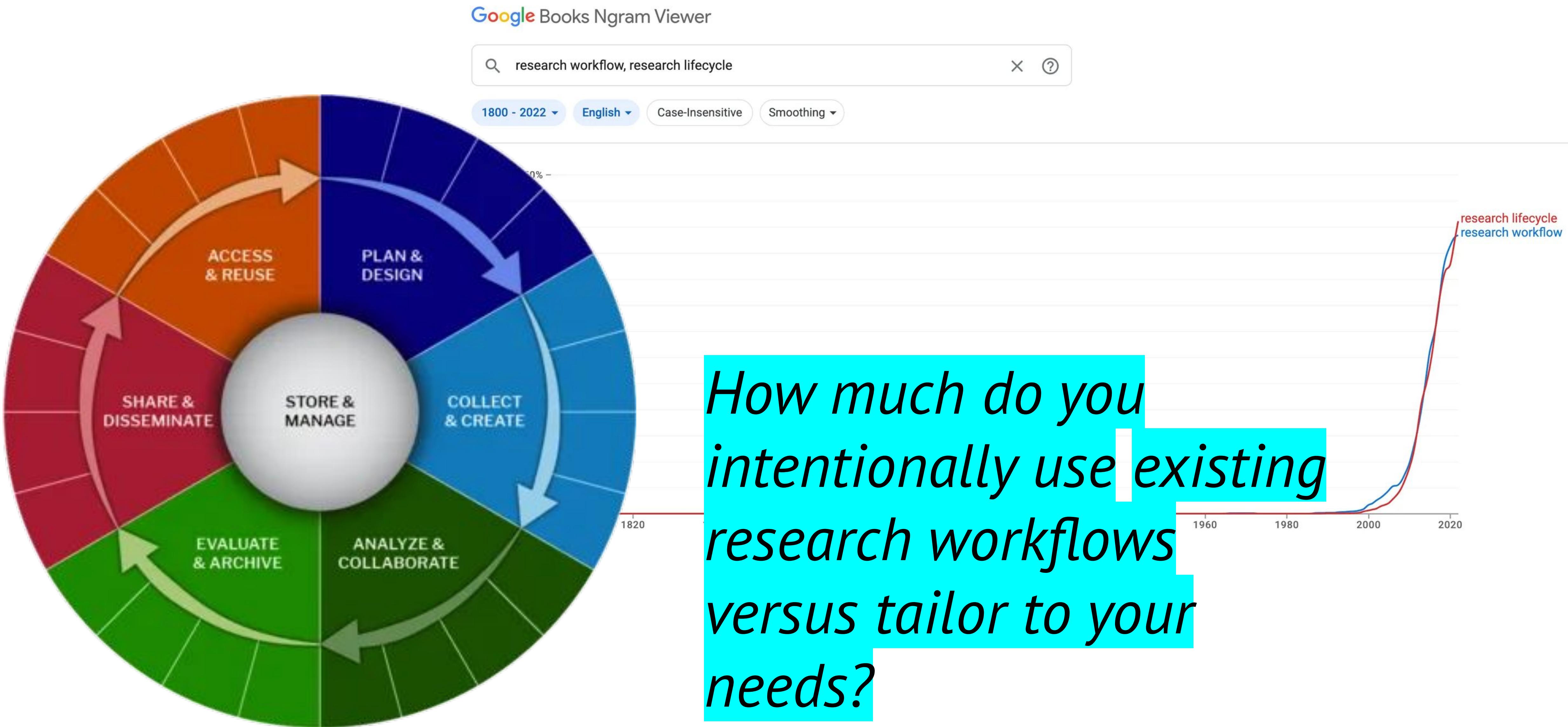
What are some the emerging trends around how academics are using AI?

ADDITIONAL SELECTED RESOURCES:

- Gefen, Alexandre, Léa Saint-Raymond, and Tommaso Venturini. "AI for Digital Humanities and Computational Social Sciences." In *Reflections on Artificial Intelligence for Humanity*, edited by Bertrand Braunschweig and Malik Ghallab, 191–202. Cham: Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-69128-8_12. **Explores how AI methodologies can bridge digital humanities and computational social sciences.**
- Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. "Can Large Language Models Transform Computational Social Science?" arXiv, February 26, 2024. <https://doi.org/10.48550/arXiv.2305.03514>. **Investigates the transformative potential of large language models in computational social science research.**
- Bail, Christopher A. "Can Generative AI Improve Social Science?" *Proceedings of the National Academy of Sciences* 121, no. 21 (May 21, 2024): e2314021121. <https://doi.org/10.1073/pnas.2314021121>. **Assesses the capacity of generative AI to advance social science research and methods.**
- Hoffmann, Manuel, Sam Boysel, Frank Nagle, Sida Peng, and Kevin Xu. "Generative AI and the Nature of Work." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, October 27, 2024. <https://doi.org/10.2139/ssrn.5007084>. **Examines how generative AI could reshape work processes and job roles, analyzing future labor implications.**

What additional research have you found useful? Do these results fit with your experiences?

Research Lifecycles & Workflows in Academia



Part 2

Brainstorming & Exploring Research Questions with AI

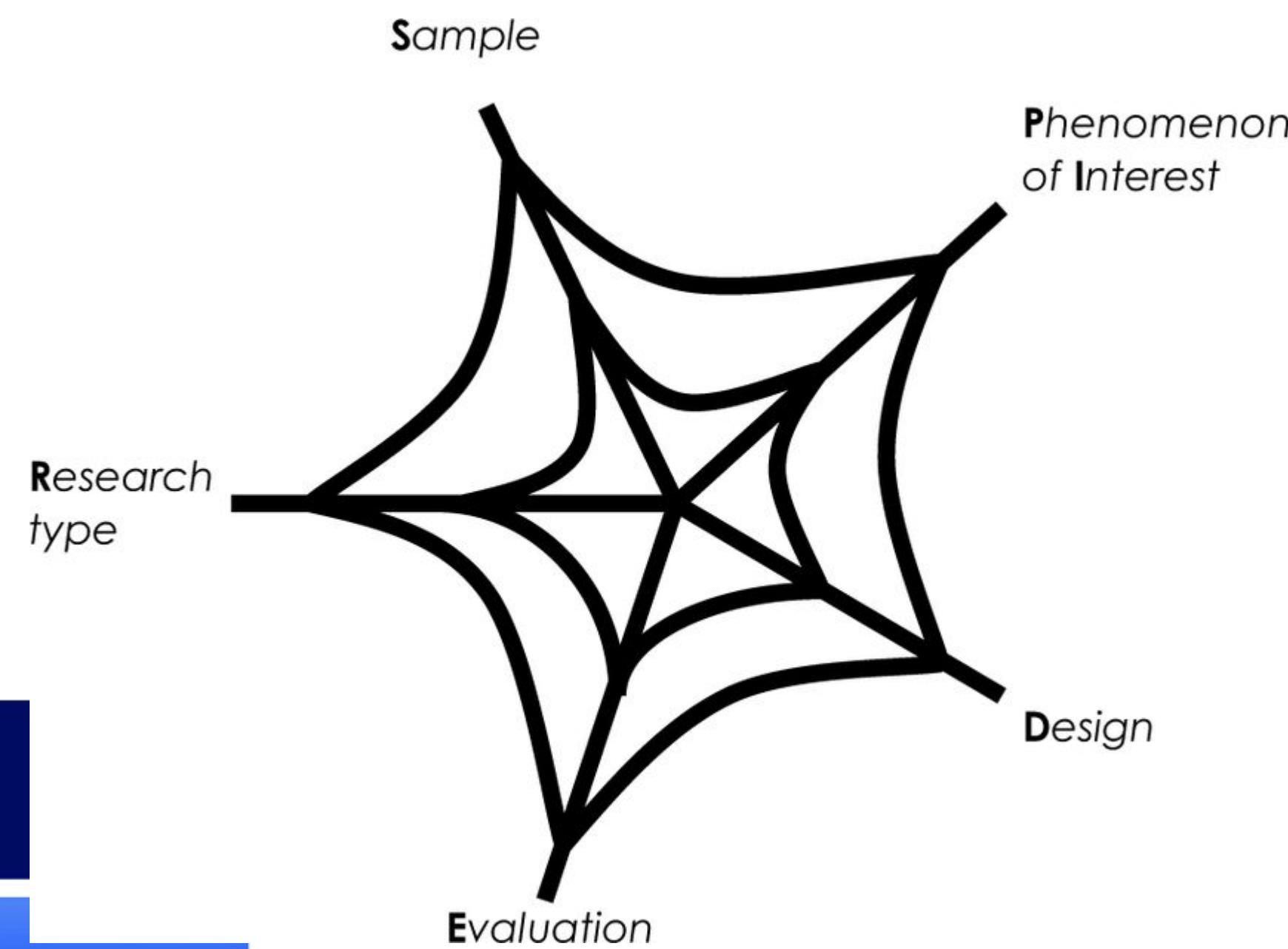
How do we begin integrating AI into our research workflows?

Brainstorming With AI

The Path to a Research Question:
From Broad Topic to a Specific Question



Lots of frameworks and approaches exist for developing research questions & topics



PICOT Framework

P	POPULATION	What is the target population?	
I	INTERVENTION	What is the intervention (treatment, diagnostic test or procedure)?	
C	COMPARISON	Is there a control group or comparison for the intervention?	
O	OUTCOME	What are the outcomes of the study?	
T	TIME	What is the time frame for the study? What is the appropriate follow-up time to assess outcome?	

Brainstorming With AI

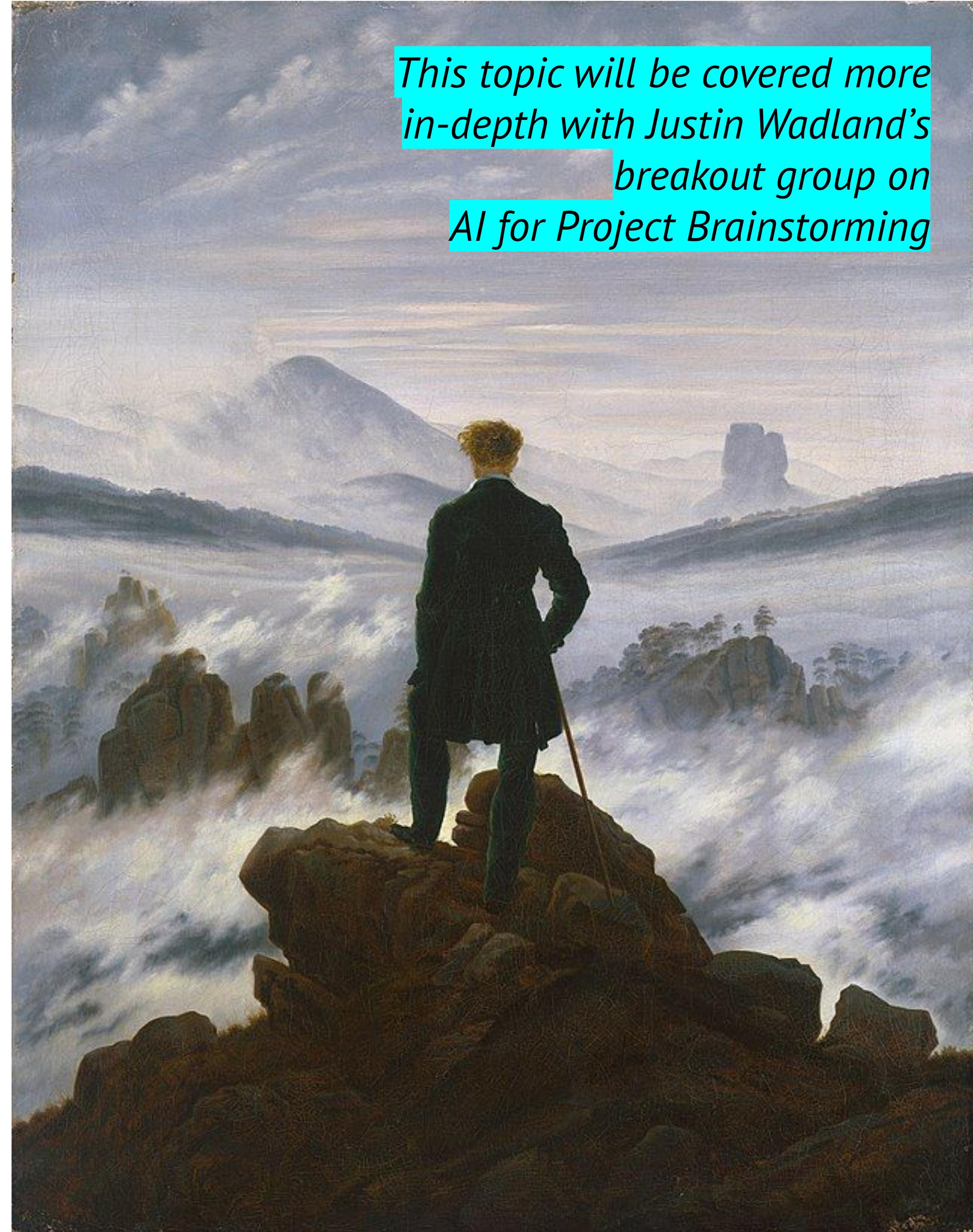
Today focusing on two scenarios:

1. OPEN HORIZONS

- In this situation, we don't have a predefined initial research question. Instead, we might have some resources or a broad area of expertise, and we are looking for suggestions on how to proceed without giving AI as much guidance or constraints, so we are using it to help us navigate the open horizon of possibilities.

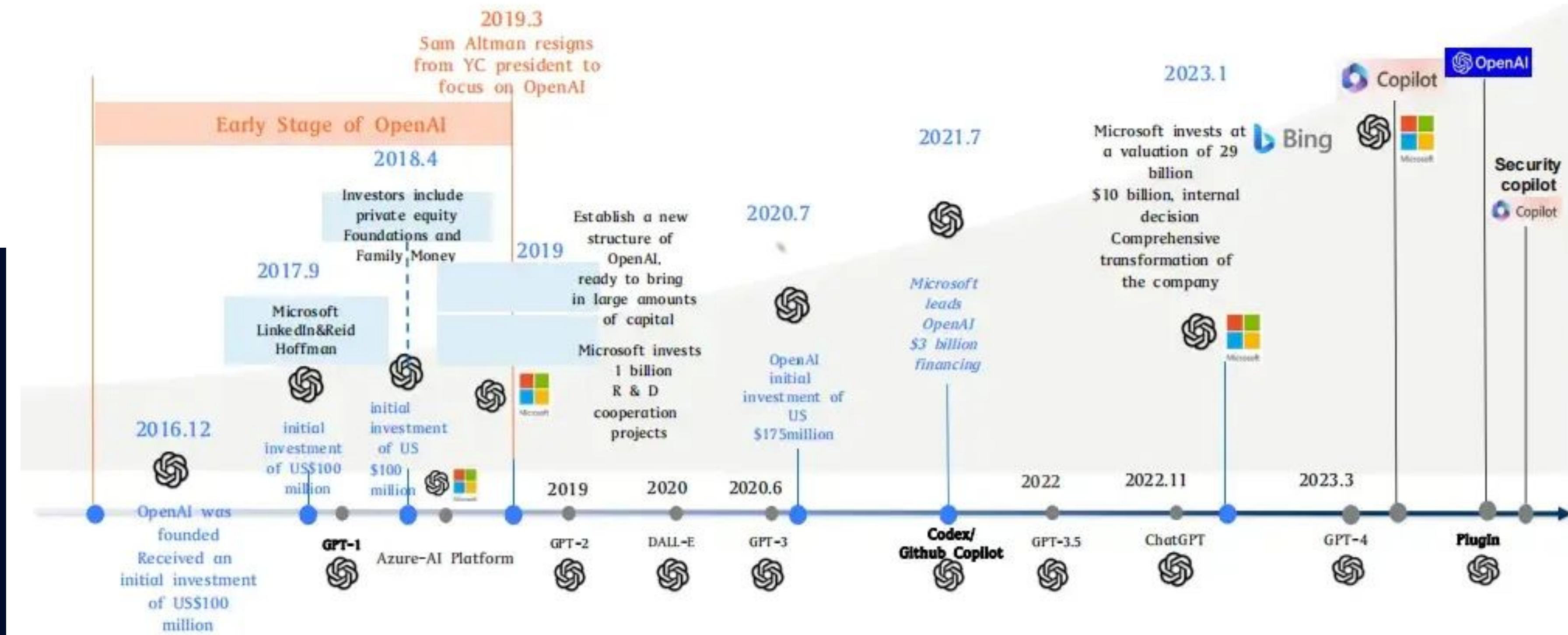
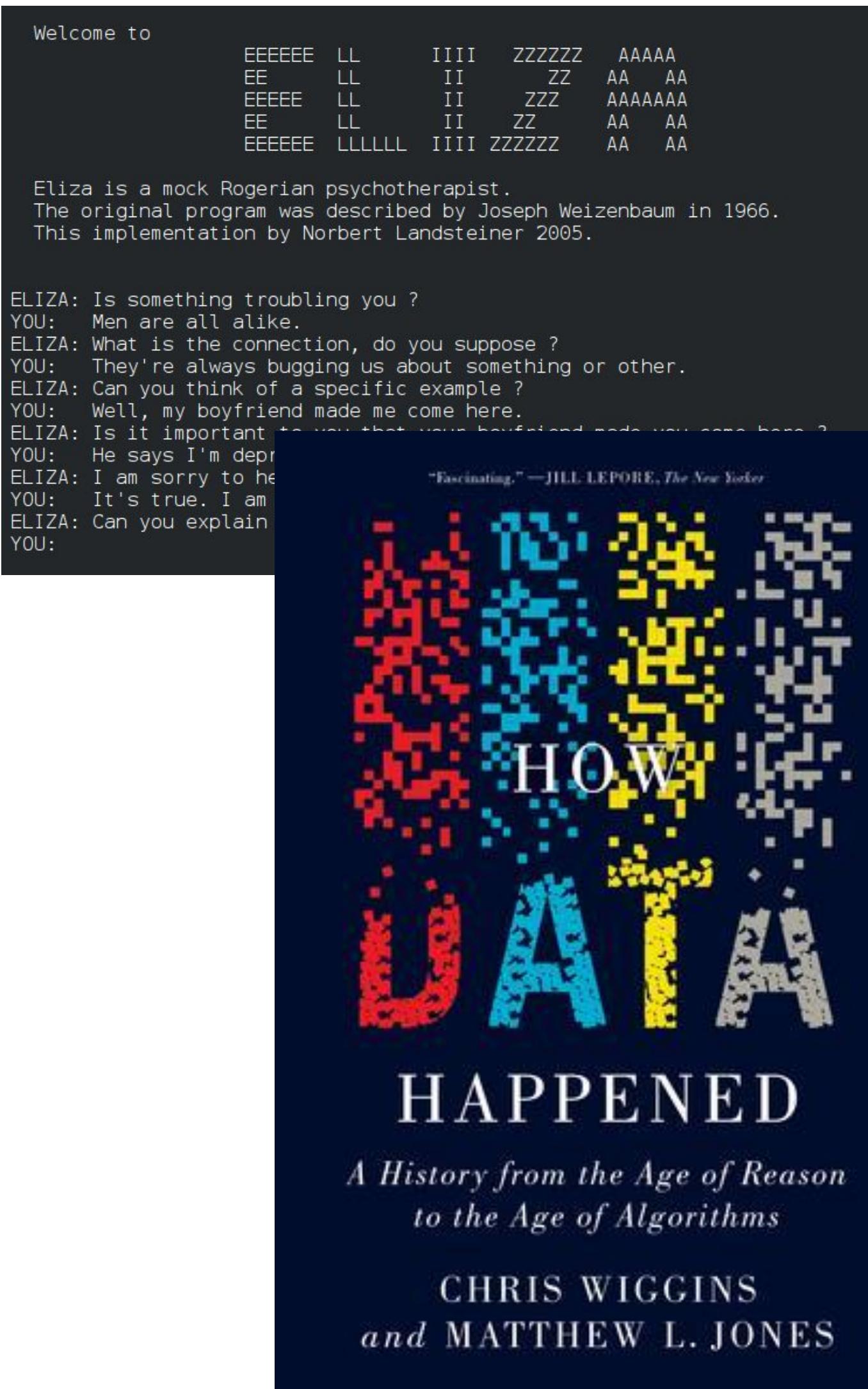
2. BLANK PAGE

- In this situation, we might have an inkling of a research question or topic, but are likely very new to the space and so are looking for guidance on how to proceed but within an area we have already defined. We are using the chatbot to essentially fill in the blank page for us and use that as a starting point for refining our research question.



This topic will be covered more in-depth with Justin Wadland's breakout group on AI for Project Brainstorming

How to Start Using AI? The Rise of the Chatbots



The Basics of Chatbots

• AI, Chatbots, and LLMs Are Distinct

- AI is a broad field, LLMs (**Large Language Models**) are a type of AI, and chatbots are just one way to interact with LLMs.
- **ChatGPT** is a chatbot that uses OpenAI's LLMs, but LLMs can also be used in search engines, coding tools, and more.

• Chatbots Are Interfaces, Not the AI Itself

- They provide a way to access LLMs, but the model exists separately. You can use the same GPT-4 model via ChatGPT, Microsoft Copilot, or even a local AI setup.

• Chatbots Are Becoming Multi-Modal

- Many now accept not just text but also images, audio, and even video as inputs.

• Chatbot Models Can Be Swapped

- Some chatbots let you choose between different models, which affects speed, accuracy, and capabilities.

The screenshot shows the ChatGPT 4.0 web interface. At the top, there are window control buttons (red, yellow, green) and a title bar with "ChatGPT 4.0 > SELECTED CHATBOT & MODEL". Below the title bar, a message from a user asks, "What can you tell me about the origins of chatbots and artificial intelligence? PROMPT FROM USER". A "MODEL RESPONSE" section follows, featuring a small circular icon with a swirl pattern. The response text reads: "The origins of chatbots and artificial intelligence (AI) are deeply intertwined with the broader history of computing, linguistics, and cognitive science. Here's a breakdown of key moments that shaped their development:"

1. Early AI Concepts (Pre-1950s)

- The idea of intelligent machines can be traced back to ancient mythology, but in the modern era, it was formalized by **Alan Turing** in his 1950 paper *Computing Machinery and Intelligence*. He proposed the famous **Turing Test**, which suggested that if a machine could hold a conversation indistinguishable from a human, it could be considered intelligent.

2. Early Chatbot-Like Programs (1950s-1970s)

- **ELIZA (1966)**: Created by **Joseph Weizenbaum** at MIT, ELIZA was one of the first chatbots. It mimicked a Rogerian psychotherapist, using pattern-matching and scripted responses to engage users.
- **PARRY (1972)**: Developed by **Kenneth Colby**, PARRY simulated a person with paranoid schizophrenia. Unlike ELIZA, it had a more defined personality and attempted to model mental illness.

How to start “prompting” with chatbots?

Visual Summary Prompt engineering Guide of OpenAI

Getting started with AI: Good enough prompting

Don't make this hard



ETHAN MOLICK
NOV 24, 2024

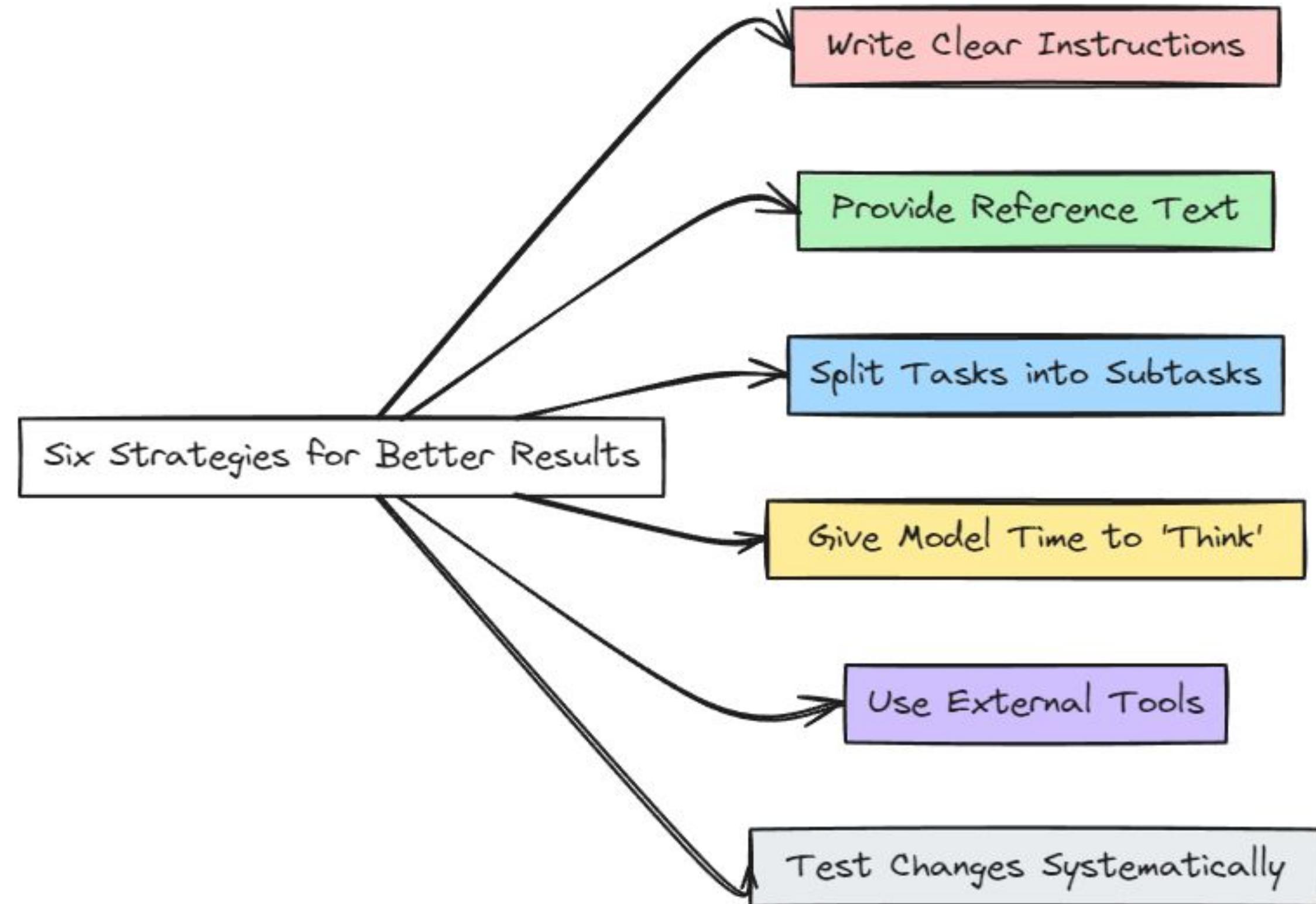


8

Good Enough Task Prompting

One of the most useful ways to use AI is to help get things done. In my book, I talk about bringing the AI to the table, trying it out for all of your work tasks to see how well it does. I still think this is the right way to start. Often, you are told to do this by treating AI like an intern. In retrospect, however, I think that this particular analogy ends up making people use AI in very constrained ways. To put it bluntly, any recent frontier model (by which I mean Claude 3.5, ChatGPT-4o, Grok 2, Llama 3.1, or Gemini Pro 1.5) is likely much better than any intern you would hire, but also weirder.

Instead, let me propose a new analogy: treat AI like an infinitely patient new coworker who forgets everything you tell them each new conversation, one that comes highly recommended but whose actual abilities are not that clear. And I mean literally *treat AI just like an infinitely patient new coworker who forgets everything you tell them each new conversation*. Two parts of this are analogous to working with humans (being new on the job and being a coworker) and two of them are very alien (forgetting everything and being infinitely patient). We should start with where AIs are closest to humans, because that is the key to good-enough prompting



How to start “prompting” with chatbots?

I'm interested in starting a new project specifically working with materials from Michigan State University's Libraries Digital Collections. Could you recommend any relevant materials in the collections considering my interests in histories of information and digital humanities?

OPEN HORIZON APPROACH

Revised “prompting” with chatbots

I'm a researcher who specializes in histories of information in the post-1945 era and a digital humanist interested in using computation to study the past. I'm interested in starting a new project specifically working with materials from Michigan State University's Libraries Digital Collections. Could you recommend any relevant materials in the collections considering my interests? Please be as specific as possible.

OPEN HORIZON APPROACH

Resources & Best Practices for Prompting

PROMPT ENGINEERING

Prompt engineering overview

Before prompt engineering

This guide assumes that you have:

1. A clear definition of the success criteria for your use case
2. Some ways to empirically test against those criteria
3. A first draft prompt you want to improve

If not, we highly suggest you spend time establishing that first. Check out [Define your success criteria](#) and [Create strong empirical evaluations](#) for tips and guidance.



Prompt generator

Don't have a first draft prompt? Try the prompt generator in the Anthropic Console!

How might I improve my prompt? How should we evaluate the results?



HOW TO PROMPT FRIENDS AND INFLUENCE PEOPLE —

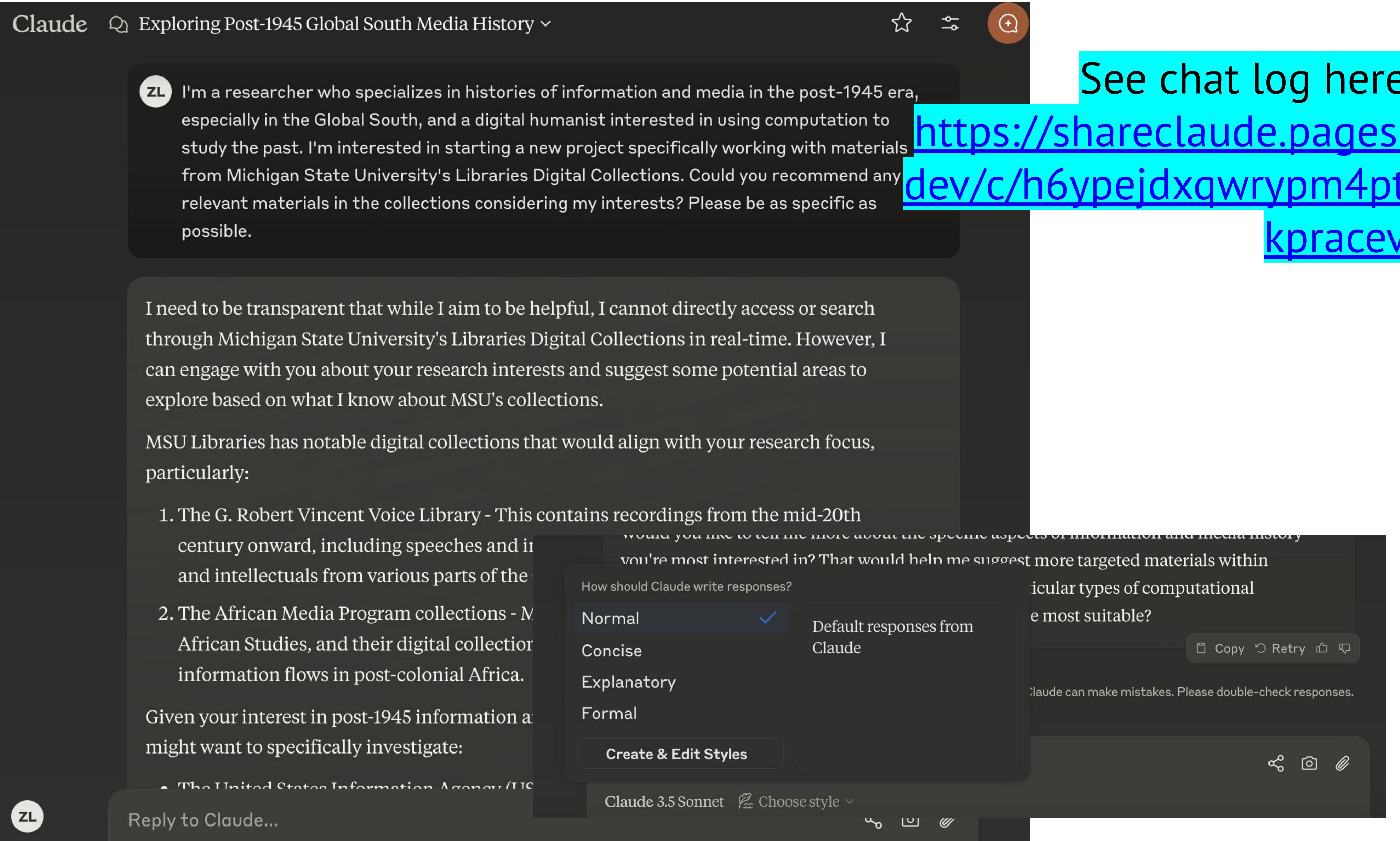
The fine art of human prompt engineering: How to talk to a person like ChatGPT

People are more like AI language models than you might think. Here are some prompting tips.

BENJ EDWARDS - 4/3/2024, 7:30 AM

Initial Results With Anthropic's Claude

 Claude
3.5 Sonnet



The screenshot shows a dark-themed interface for Claude 3.5 Sonnet. At the top, it says "Claude Exploring Post-1945 Global South Media History". On the right, there are icons for star, share, and a plus sign.

User Message:

I'm a researcher who specializes in histories of information and media in the post-1945 era, especially in the Global South, and a digital humanist interested in using computation to study the past. I'm interested in starting a new project specifically working with materials from Michigan State University's Libraries Digital Collections. Could you recommend any relevant materials in the collections considering my interests? Please be as specific as possible.

Claude Response:

I need to be transparent that while I aim to be helpful, I cannot directly access or search through Michigan State University's Libraries Digital Collections in real-time. However, I can engage with you about your research interests and suggest some potential areas to explore based on what I know about MSU's collections.

MSU Libraries has notable digital collections that would align with your research focus, particularly:

1. The G. Robert Vincent Voice Library - This contains recordings from the mid-20th century onward, including speeches and interviews with political leaders and intellectuals from various parts of the world.
2. The African Media Program collections - These collections focus on African Studies, and their digital collections document the flow of information in post-colonial Africa.

Given your interest in post-1945 information and media history, you might want to specifically investigate:

Would you like to tell me more about the specific aspects of information and media history you're most interested in? That would help me suggest more targeted materials within particular types of computational models most suitable?

How should Claude write responses?
 Normal
 Concise
 Explanatory
 Formal

Create & Edit Styles

Default responses from Claude

Claude can make mistakes. Please double-check responses.

Reply to Claude...

Choose style

See chat log here
<https://shareclaude.pages.dev/c/h6ypejdxqwrypm4pt>
kpracev

Distinguishing Between Chatbots

Service	Best Model	Live Mode	"Reasoning"	Web Access	Generates Images	Deep Research	Executes Code	Data Analysis	Sees images	Sees video	Reads Docs	Personality	Superpower
OpenAI ChatGPT	GPT-4o	✓ Full multimodal	✗	✓	✓ DALL-E3	✓	✓	✓	✓	In Live Mode	✓	Polished and efficient in text. In live mode, expressive and adaptive.	Live mode, most versatile set of features and capabilities
	o1/o3 family	✗	✓	✗	✗	✓	✓	✗	✓	✗	✗	Methodical and analytical	Very powerful model for complex reasoning tasks, particularly in science, coding, and mathematics
Microsoft Copilot	"Copilot"	Voice only	✓	✓	✓ DALL-E3	✗	Limited	✗	✓	✗	✓	Since it uses different models behind the scenes, a little inconsistent	Works well with Microsoft products and services
Anthropic Claude	Claude 3.5	✗	✗	✗	✗	✗	✓	Limited	✓	✗	✓	Clever and friendly	Often the most creative and socially engaging model
Google Gemini	Gemini family	Voice only	✓	✓	✓ Imogen-3	✓	Limited	Limited	✓	✓	✓	Helpful and a bit bland	Wide variety of features, good connections with search
X.ai Grok	Grok-2	✗	✗	✓ Mostly X	✓ Aurora	✗	✗	✗	✓	✗	✓	Sarcastic and "fun" (though you can tone that down)	Powerful model integrated tightly with X
DeepSeek	DeepSeek v3	✗	✓	✓	✗	✗	✗	✗	✓	✗	Limited	Neurotically helpful, warm	Remarkably cheap and powerful model out of China

Which AI to Use Now: An Updated Opinionated Guide (Updated Again 2/15)

Picking your general-purpose AI



ETHAN MOLICK
JAN 26, 2025

This topic will be covered more in-depth with Yuri Cantrell's breakout group on Identifying AI Tools for Research

Core Concepts With AI: Training Data



Common Crawl
maintains a [free, open repository](#) of web crawl data that can be used by anyone.

Common Crawl is a 501(c)(3) non-profit founded in 2007.
We make wholesale extraction, transformation and analysis of open web data accessible to researchers.

Over 250 billion pages spanning 17 years.
Free and open corpus since 2007.
Cited in over 10,000 research papers.
3–5 billion new pages added each month.

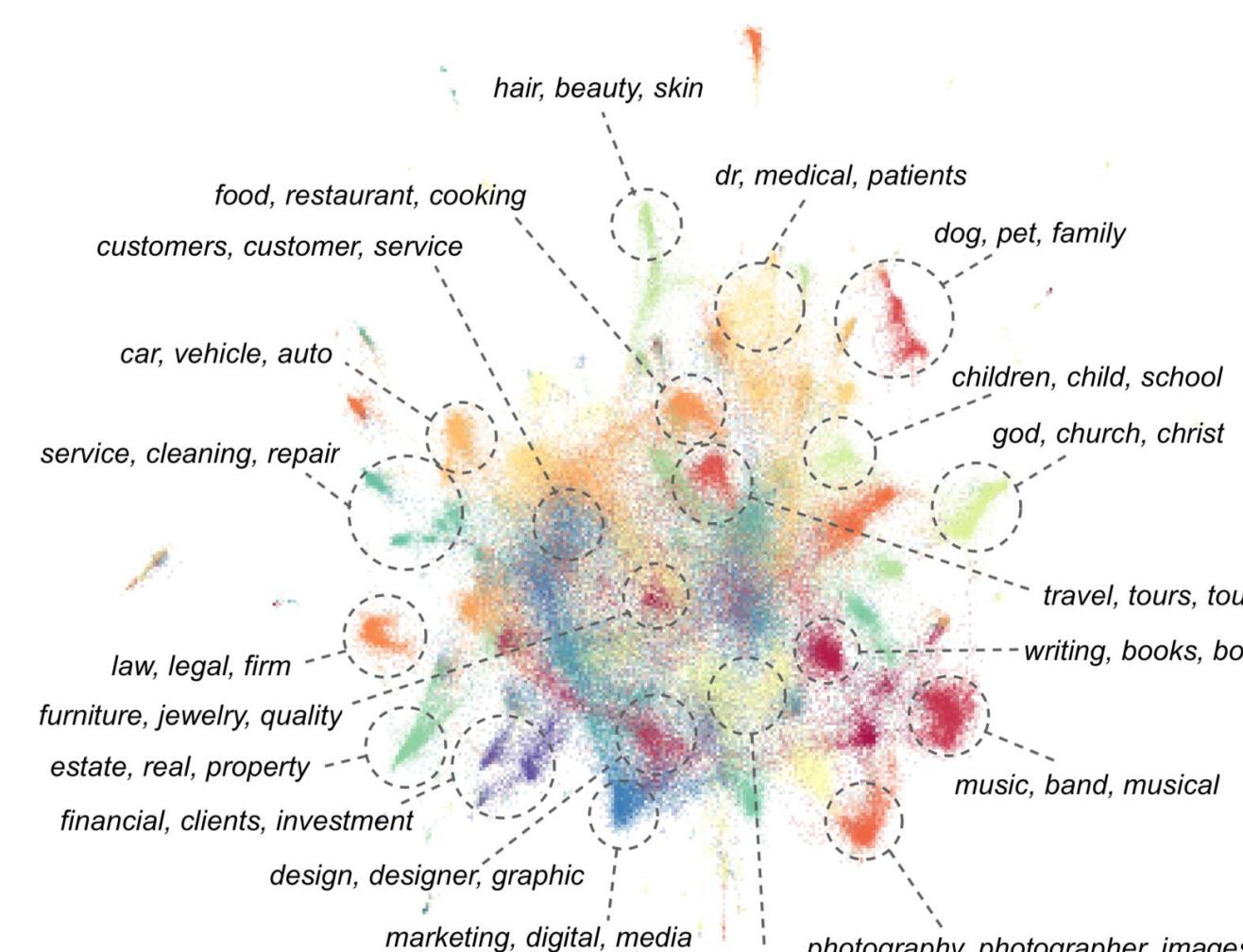


LDNOOBW/List-of-Dirty-Naughty-Obscene-and...

List of Dirty, Naughty, Obscene, and Otherwise Bad Words



Inside the secret list of websites that make AI like ChatGPT sound smart



AboutMe: Using Self-Descriptions in Webpages to Document the Effects of English Pretraining Data Filters

Li Lucy^{1,2} Suchin Gururangan⁵ Luca Soldaini¹
Emma Strubell^{1,4} David Bamman² Lauren Klein³ Jesse Dodge¹
¹Allen Institute for AI ²University of California, Berkeley ³Emory University
⁴Carnegie Mellon University ⁵University of Washington
lucy3_li@berkeley.edu

Abstract

Large language models' (LLMs) abilities are drawn from their pretraining data, and model development begins with data curation. However, decisions around what data is retained or removed during this initial stage is underscrutinized. In our work, we ground web text, which is a popular pretraining data source, to its social and geographic contexts. We create a new dataset of 10.3 million self-descriptions of

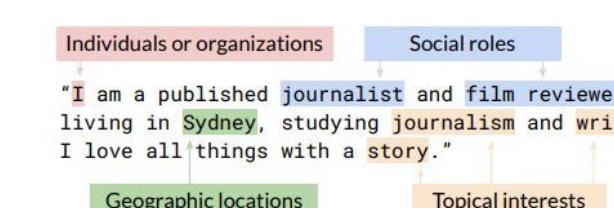
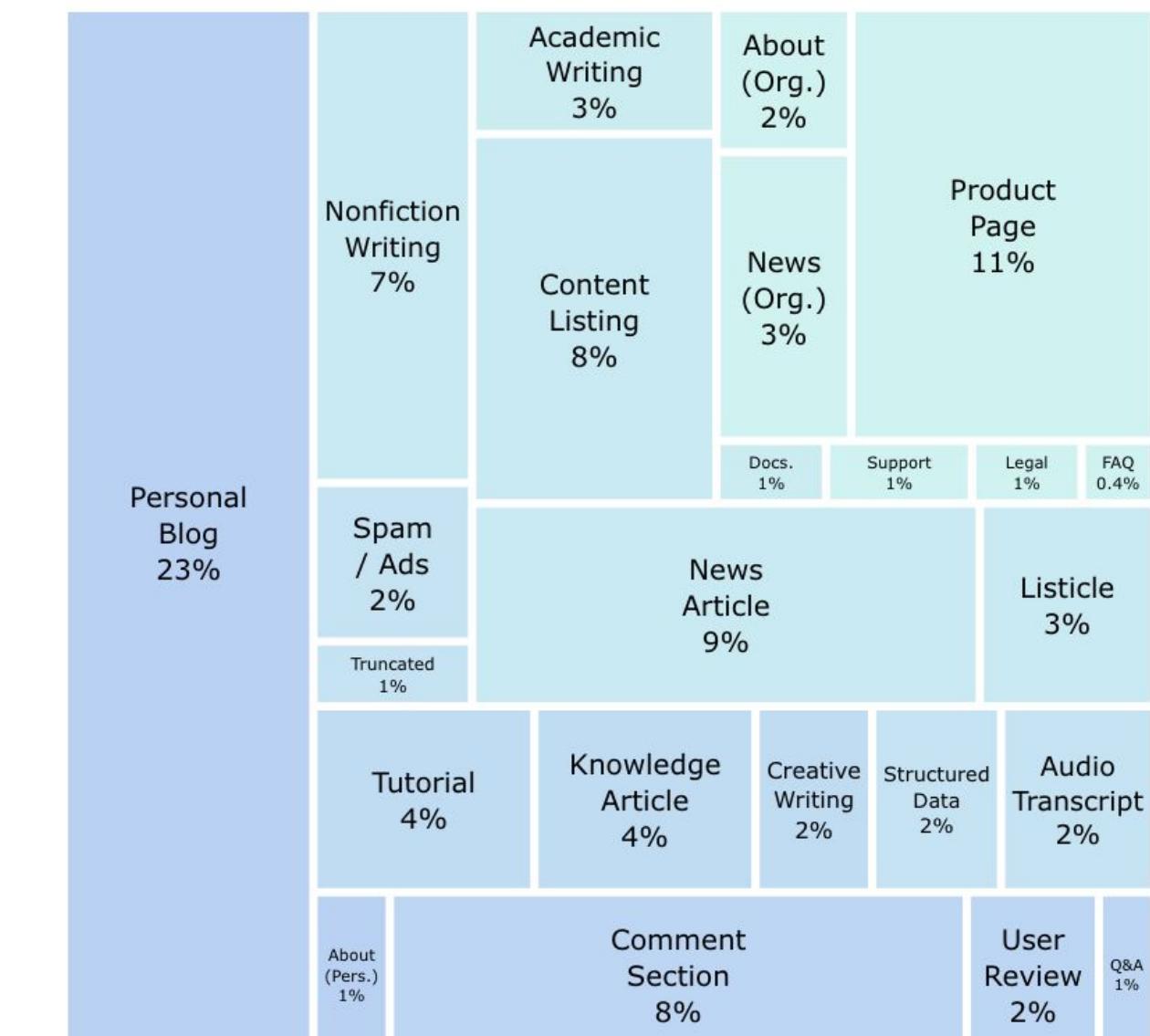
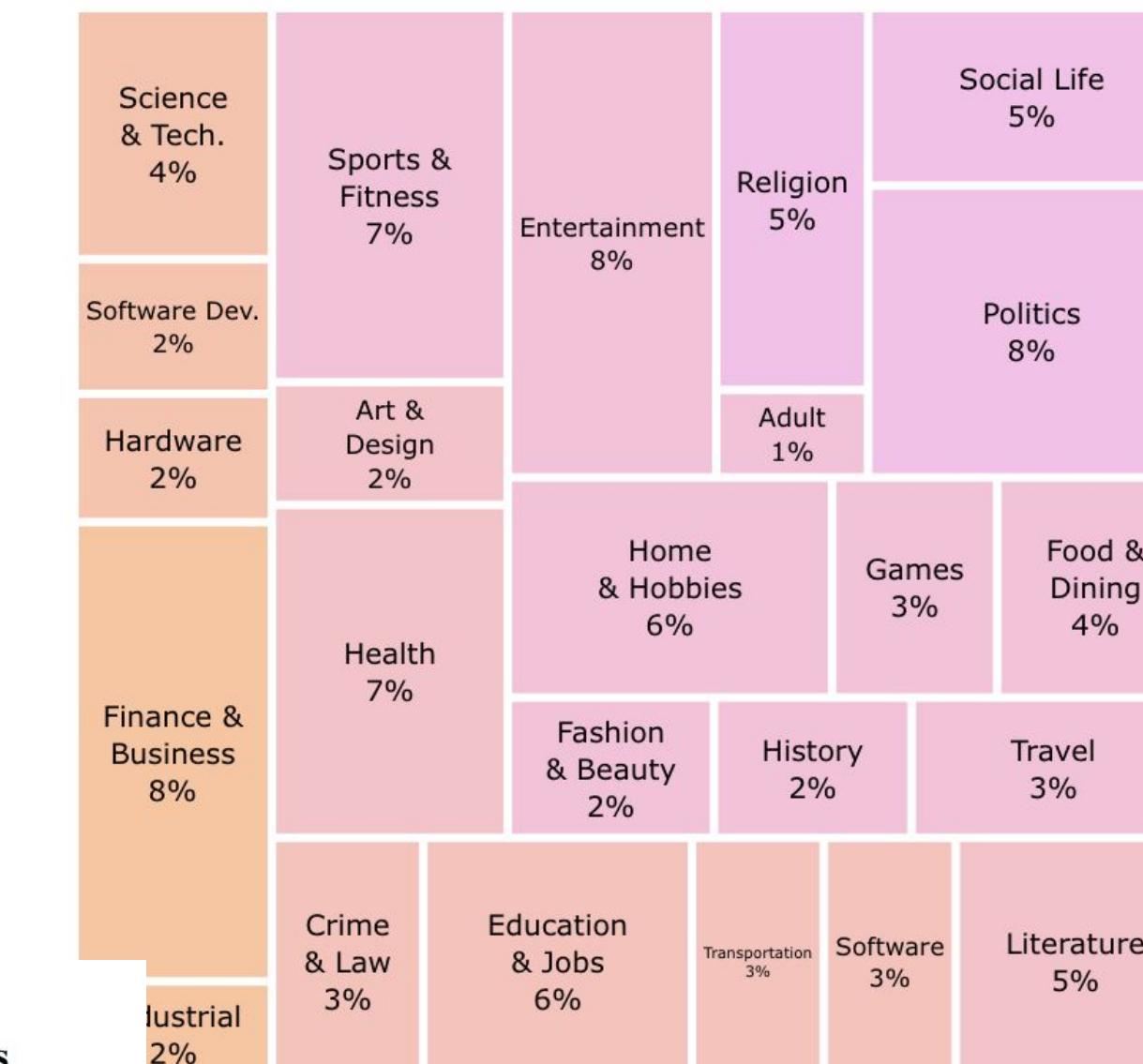


Figure 1: A paraphrased excerpt from a website's ABOUT page, with commonly stated social dimensions emphasized with highlighting.

Explore the Corpus

Explore different topics and formats and click to see examples! We visualize a [pre-training corpus](#) derived from CommonCrawl.
The areas show the number of tokens per domain. Move your mouse over a topic to see the distribution of formats for that topic or vice versa.



Organize the Web: Constructing Domains Enhances Pre-Training Data Curation

Alexander Wetting^{1,2} Kyle Lo² Sewon Min^{3,2} Hannaneh Hajishirzi^{2,4} Danqi Chen¹ Luca Soldaini²

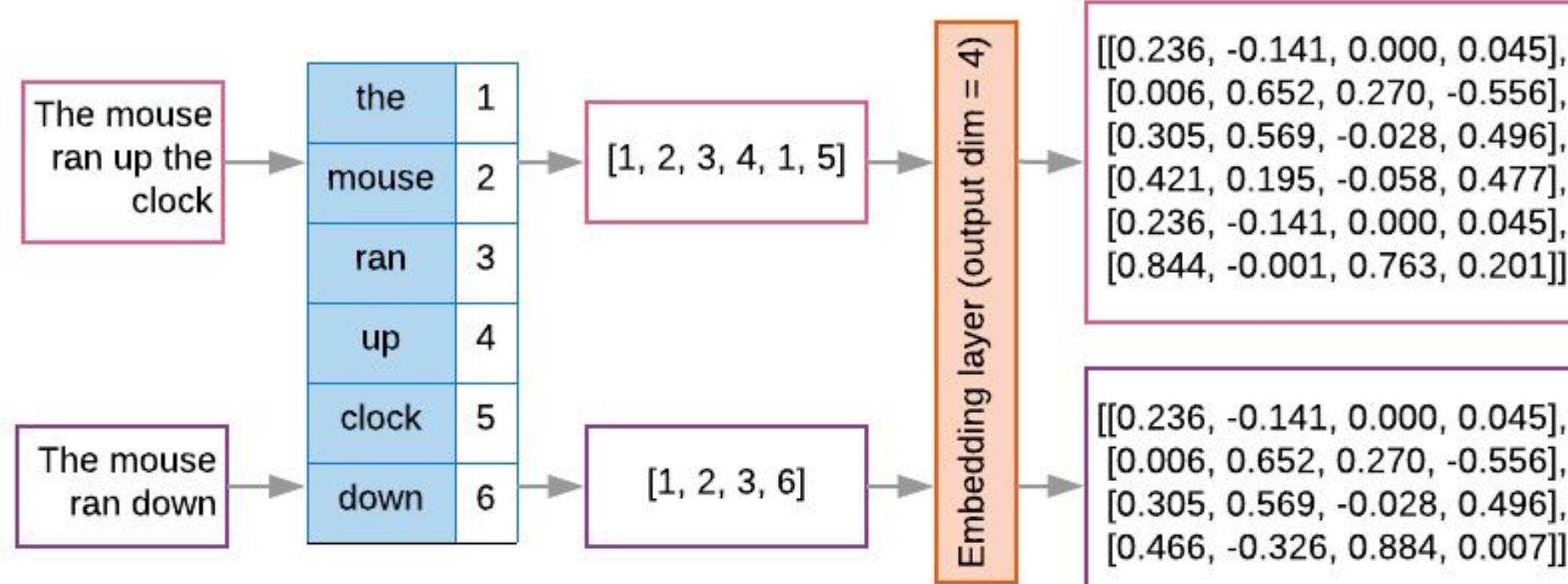
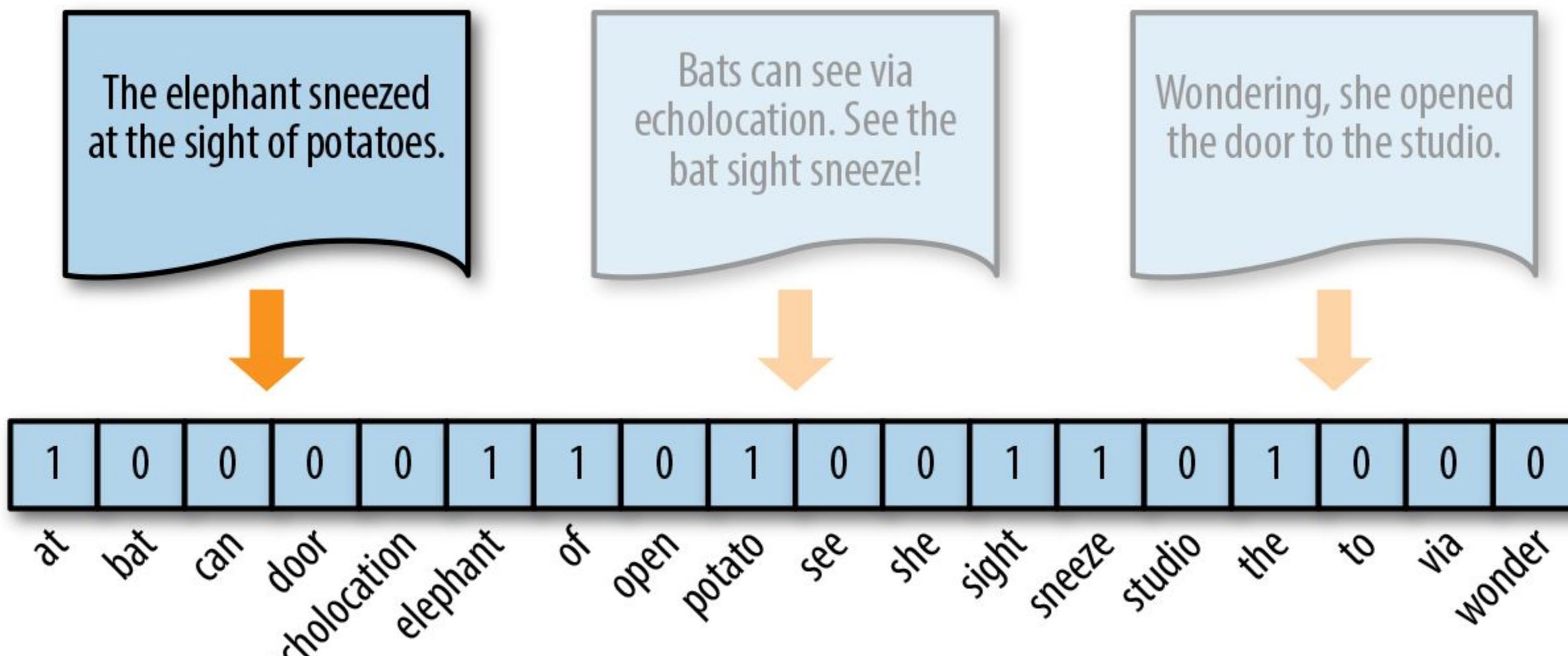
¹Princeton University ²AI2 ³UC Berkeley ⁴University of Washington

Paper

Hugging Face

Code

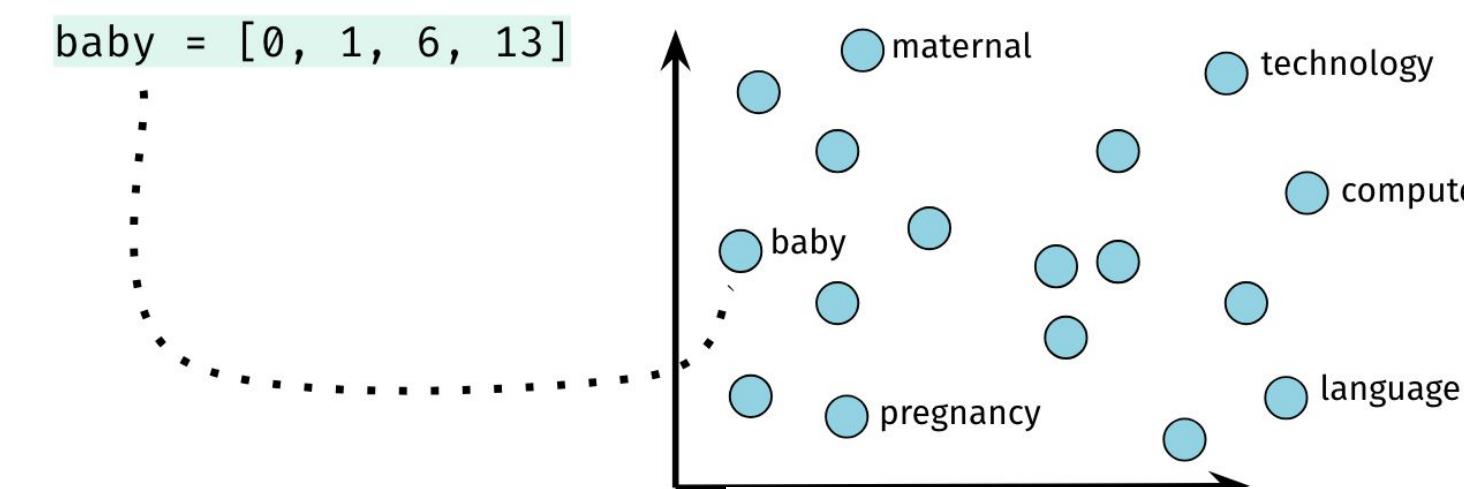
Core Concepts With AI: Tokenization & Vectorization



Measure word similarity

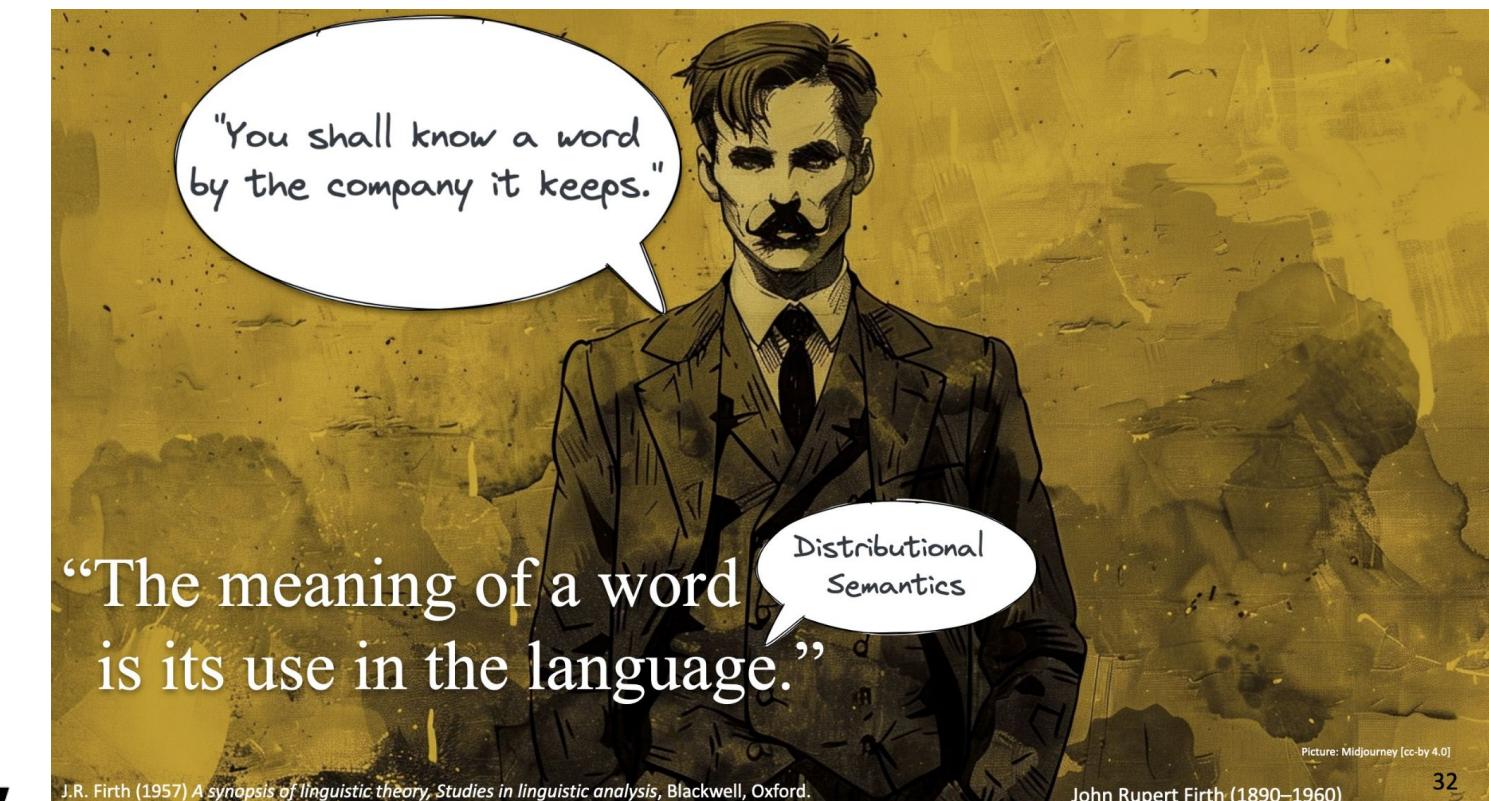
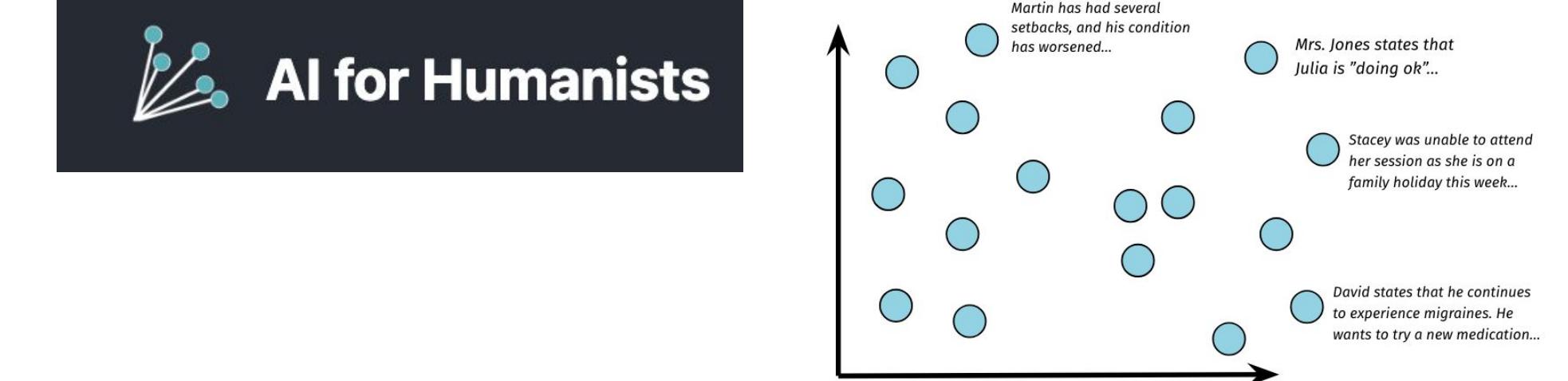
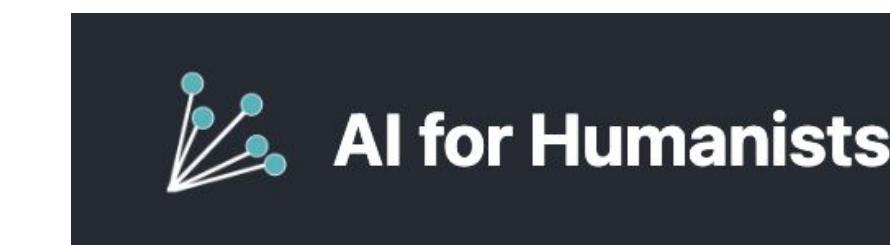
We can plot the words in a graph.

And then we can measure distances and relationships between words!



Measure document similarity

For whole documents, we can **average** all the word vectors to get a single vector for the document. Now we can measure distances and relationships between the documents.



Distinguishing Between Chatbots: Costs & Privacy

- **Most Chatbots Operate as SaaS Platforms**

- Chatbots are typically **Software as a Service (SaaS) platforms**, meaning they require an internet connection and operate via a cloud-based model.
- Many offer **free tiers**, but these often come with limitations, such as access to lower-tier models, restricted features (e.g., file uploads, extended memory), or usage caps. **Paid plans** provide better models and advanced capabilities (e.g., GPT-4o, Claude 2, Gemini Advanced).

- **Understanding Queries vs. Tokens**

- Queries refer to each prompt or request you send to the chatbot. Free and paid plans often **limit the number of queries per time period**.
- Tokens measure how much text is processed—both in your prompt and the AI's response.
- **APIs charge based on tokens**, making pricing less predictable since tokenization varies across companies, languages, and models.

The screenshot shows the OpenAI pricing page with three plan options:

- Free**: \$0 USD/month. Description: "Explore how AI can help you with everyday tasks". Call-to-action: "Your current plan". Features: ✓ Access to GPT-4o mini, ✓ Standard voice mode, ✓ Limited access to GPT-4o, ✓ Limited access to file uploads, advanced data analysis, web browsing, and image generation, ✓ Use custom GPTs.
- Plus (POPULAR)**: \$20 USD/month. Description: "Level up productivity and creativity with expanded access". Call-to-action: "Get Plus". Features: ✓ Everything in Free, ✓ Extended limits on messaging, file uploads, advanced data analysis, and image generation, ✓ Standard and advanced voice mode, ✓ Limited access to o1 and o1-mini, ✓ Opportunities to test new features, ✓ Create and use custom GPTs.
- Pro**: \$200 USD/month. Description: "Get the best of OpenAI with the highest level of access". Call-to-action: "Get Pro". Features: ✓ Everything in Plus, ✓ Unlimited access to o1, o1-mini, and GPT-4o, ✓ Unlimited access to advanced voice, ✓ Access to o1 pro mode, which uses more compute for the best answers to the hardest questions.

ANTHROPIC

Model Pricing

Model family	Best for	Context window	Standard Pricing
Claude Instant	Low latency, high throughput use cases	100,000 tokens	Prompt \$1.63/million tokens Completion \$5.51/million tokens
Claude 2	Superior performance on tasks that require complex reasoning Claude 2 is Anthropic's best-in-class offering	100,000 tokens	Prompt \$11.02/million tokens Completion \$32.68/million tokens

*Claude 1 is still accessible and offered at the same price as Claude 2.

Initial Results With Microsoft's Co-Pilot

The screenshot displays the Microsoft Copilot interface, showing a conversation between a user and the AI. The user's message is highlighted in blue, while the AI's responses are in white.

User Message:

I'm a researcher who specializes in histories of information and media in the post-1945 era, especially in the Global South, and a digital humanist interested in using computation to study the past. I'm interested in starting a new project specifically working with materials from Michigan State University's Libraries Digital Collections. Could you recommend any relevant materials in the collections considering my interests? Please be as specific as possible.

Copilot Response:

You can find your page here: [I'm a researcher who](#). If you want to add other responses to it, just click "Edit in Pages".

AI-generated Content (List 1):

- African e-Journals Project:** This collection includes digitized journals published in Africa, covering a wide range of topics including media, communication, and information studies. It could provide valuable insights into post-1945 media landscapes in the Global South [1](#).
- East African Newspapers:** This collection features newspapers from Kenya, Uganda, and Tanzania, dating from the 1940s to the 1970s. These primary sources can offer a rich perspective on the dissemination of information and media practices in the region during the post-1945 era [1](#).

Ask me anything...

0 / 128000

AI-generated Content (List 2):

digitized materials, including documents, newspapers, journals, images, and data. You can search for specific topics related to your research interests and find relevant materials [1](#).

These collections should provide a solid foundation for your project. If you need more specific guidance or assistance, the MSU Libraries' Digital Repository Team can help you navigate the collections and find the materials most relevant to your research [1](#).

Good luck with your project! If you have any other questions or need further assistance, feel free to ask.

Add response to an editable page in the right pane

[Edit in Pages](#) [Copy](#) [Like](#) [Unlike](#) AI-generated content may be incorrect

Copilot Response:

You can find your page here: [I'm a researcher who](#). If you want to add other responses to it, just click "Edit in Pages".

View prompts

Ask me anything...

0 / 128000

AI-generated Content (List 3):

- African e-Journals Project:** This collection includes digitized journals published in Africa, covering a wide range of topics including media, communication, and information studies. It could provide valuable insights into post-1945 media landscapes in the Global South [1](#).
- East African Newspapers:** This collection features newspapers from Kenya, Uganda, and Tanzania, dating from the 1940s to the 1970s. These primary sources can offer a rich perspective on the dissemination of information and media practices in the region during the post-1945 era [1](#).
- African Activist Archive:** This archive documents the history of activism in Africa, including materials related to media and information dissemination. It includes posters, pamphlets, and other media that can shed light on the role of information in social movements [1](#).
- Latino and Latin American Collections:** These collections include a variety of materials such as newspapers, magazines, and other media from Latin America. They can provide insights into the media history and information practices in the Global South, particularly in the post-1945 period [1](#).
- South Asian American Digital Archive (SAADA):** While focused on the South Asian diaspora in the United States, this collection includes materials that reflect the media and information practices of South Asian communities, which could be relevant to your study of the Global South [1](#).
- Digital Repository:** The MSU Libraries Digital Repository offers a wide range of

More Information:

more specific guidance or assistance, the MSU Libraries' Digital Repository Team can help you navigate the collections and find the materials most relevant to your research [1](#).

If you have any other questions or need further assistance, feel free to ask.

[Edit in Pages](#) [Copy](#) [Like](#) [Unlike](#) AI-generated content may be incorrect

Distinguishing Between Chatbots: Costs & Privacy

- **Data Leakage**
 - AI models trained on user inputs may store, recall, or **regurgitate private data**, leading to exposure in AI-generated outputs.
 - Some chatbots store interactions and use them to further train models unless explicitly disabled to create **synthetic data**.
- **Data Privacy Unknown**
 - While many chatbot providers claim they do not store or train on user data, there is often **little transparency or external verification**.
 - No clear auditing standards exist to confirm how data is handled, making it **risky for particularly sensitive or confidential information**.

NDC { Porto }

Keynote: Can you trust your (large language) model?

Jodie Burchell
Developer Advocate in Data Science

WIRED



COLLECTIVE ACTION

DeepSeek's Popular AI App Is Explicitly Sending US Data to China

Amid ongoing fears over TikTok, Chinese generative AI platform DeepSeek says it's sending heaps of US user data straight to its home country, potentially setting the stage for greater scrutiny.

MATT BURGESS AND LILY HAY NEWMAN
81.27.25 17:01 PM



Core Concepts With AI: Memorization & Proprietary Data

 Community Article

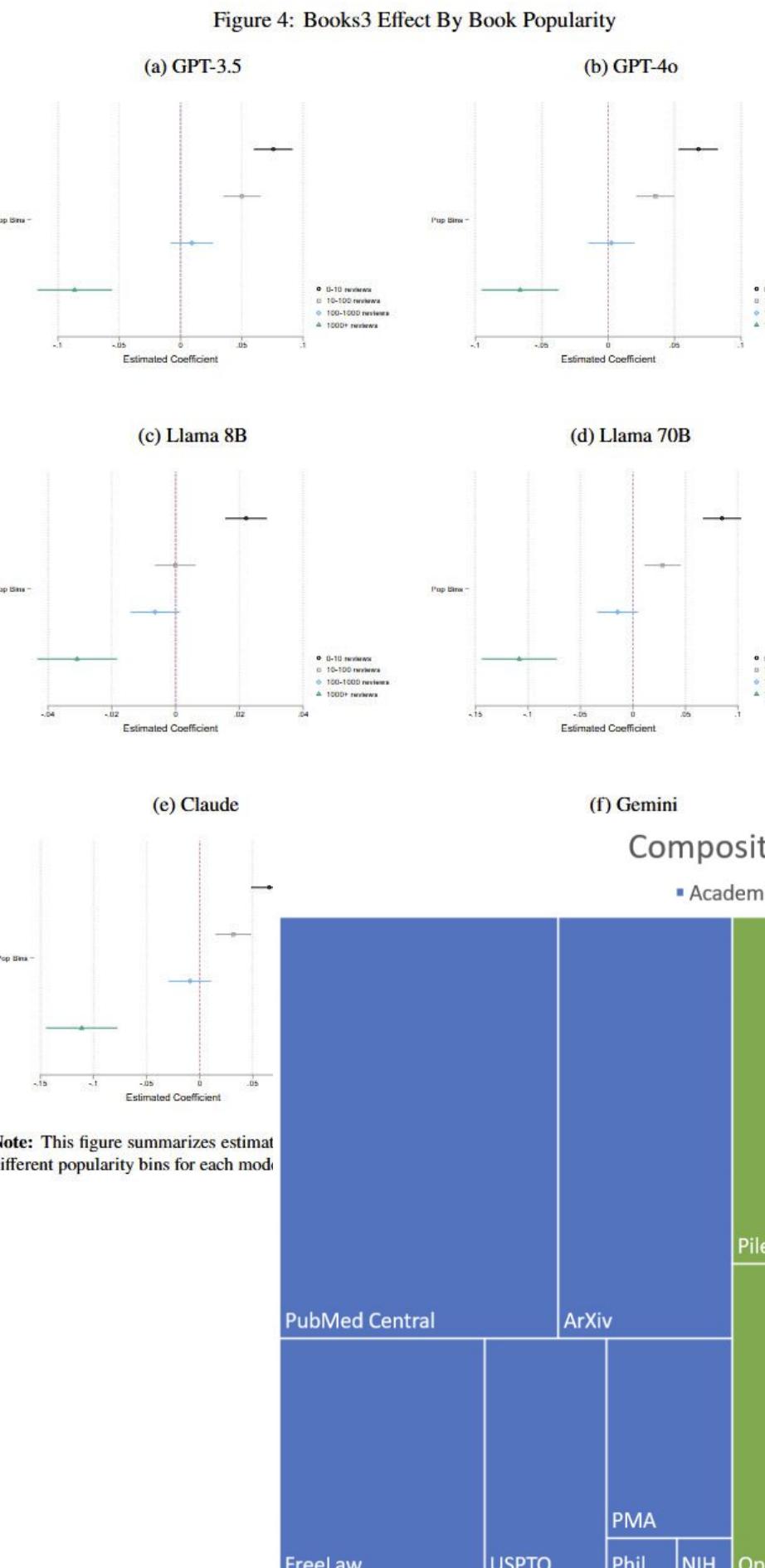


Table 1

Results of three prompting methods on poetry retrieval ability of four different language models. No. Retrieved is out of 240 total poems.

Prompt	Model	No. Retrieved
Can you please retrieve the full text of the poem [Title] by [Poet] for me?	ChatGPT	71
	PaLM	9
	Pythia	0
	GPT-2	0
Can you please retrieve the full text of the poem that begins [First Line] by [Poet] for me?	ChatGPT	72
	PaLM	10
	Pythia	0
	GPT-2	0
[First Line]		ChatGPT 26

[First Line] The Chatbot and the Canon: Poetry Memorization in LLMs

Lyra D'Souza¹, David Mimno²

¹Department of Computer Science, Cornell University, USA

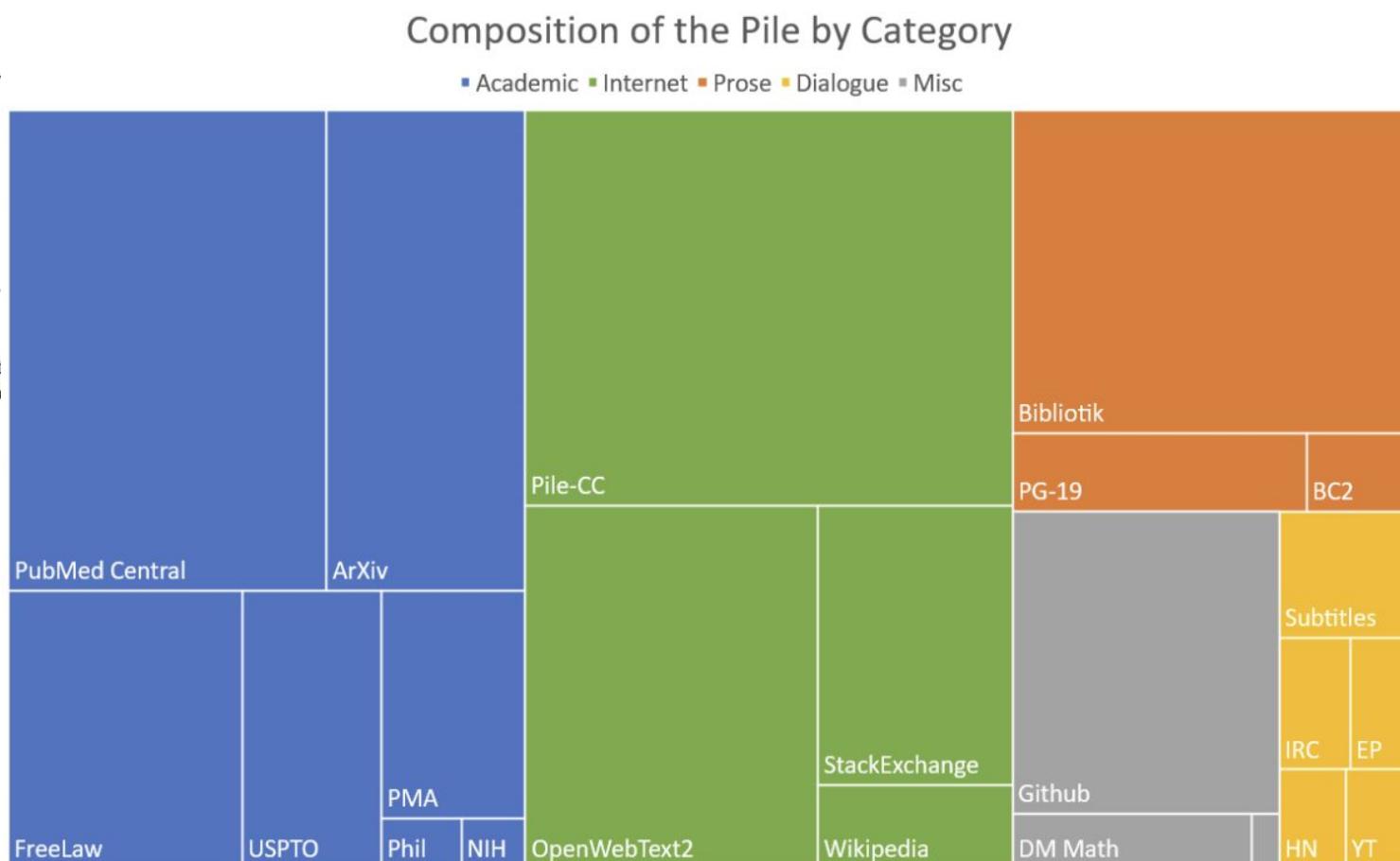
²Department of Information Science, Cornell University, USA

Abstract

Large language models are able to memorize and generate long passages of text from their pretraining data. Poetry is commonly available on the web and often fits within language model context sizes. As LLMs continue to grow as a tool in literary analysis, the accessibility of poems will determine the effective canon. We assess whether we can prompt current language models to retrieve existing poems, and what methods lead to the most successful retrieval. For the highest performing model, ChatGPT, we then evaluate which features of poets best predict memorization, as well as document changes over time in ChatGPT's ability and willingness to retrieve poetry.

Keywords

memorization, LLMs, poetry, ChatGPT, prompting, training data



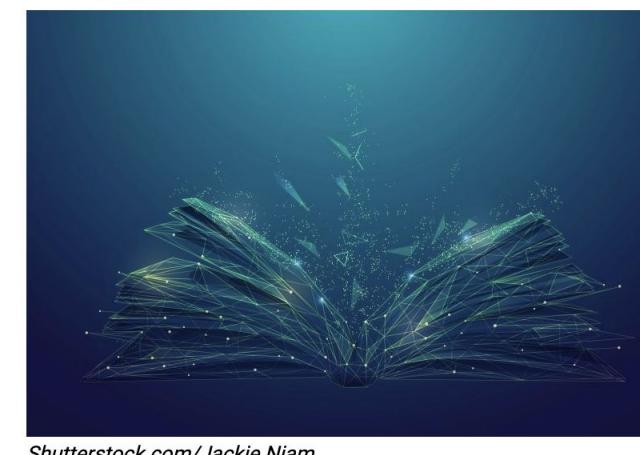
Releasing Common Corpus: the largest public domain dataset for training LLMs

by  Pclanglais

[HOME](#) / [NEWS STORIES](#) / Clarivate launches subscription-based access strategy

Clarivate launches subscription-based access strategy

18 February 2025



Shutterstock.com/Jackie Niam

Clarivate has launched a subscription-based content access strategy, while also phasing out one-time perpetual purchases of digital collections, print and digital books for libraries.

The new approach, will be introduced throughout 2025 "breaks down barriers through broad and affordable access to highly curated, trusted academic resources, including Ebooks and primary digital collections, enabling every library to support research, teaching and learning goals".

Two products are being introduced:

ProQuest Ebooks:

- More than 700,000 Ebooks, across 10 core disciplines, plus additional essential interdisciplinary titles;
- A new Ebook Central Reader app to enable students to read and bookmark on the go; and
- The addition of Ebook Central Research Assistant, a new AI tool designed to enhance student learning and streamline research.

Core Concepts With AI: Embeddings & Prediction

Thinking Like an AI

A little intuition can help



ETHAN MOLICK
OCT 20, 2024

This breaks up words (even phan t a s mag or ically long words) into token s

It was the White Rabbit, trotting slowly back again, and looking anxiously about as it went, as if it had lost something: and she heard it muttering to itself 'The Duchess! The Duchess! Oh my dear paws! Oh my fur and whiskers! She'll get me executed, as sure as ferrets are ferrets! Where can I have dropped them, I wonder?' Alice guessed in a moment that it was looking for the fan and the pair of white kid gloves, and she very good-naturedly began hunting about for them, but they were nowhere to be seen—everything seemed to have changed since her swim in the pool, and the great hall, with the glass table and the little door, had vanished completely.

Duchess = 98.55%

Duch = 1.28%

Du = 0.04%

Dutch = 0.03%

Duc = 0.02%

D = 0.01%

= 0.01%

Dou = 0.01%

Duke = 0.01%

\n\n = 0.01%

Total: -0.01 logprob on 1 tokens
(99.96% probability covered in top 10 logits)

How does Chat GPT work?

Training Data

More Sentences

"Early bird gets the worm."

Ask Chat GPT to predict (generate) the next word.

"Early bird gets the ???

Transformer



worm.

Reward the model if it predicts worm...

...and we repeat for every sentence we can find!

Generative

We are asking Chat GPT to generate the next word!

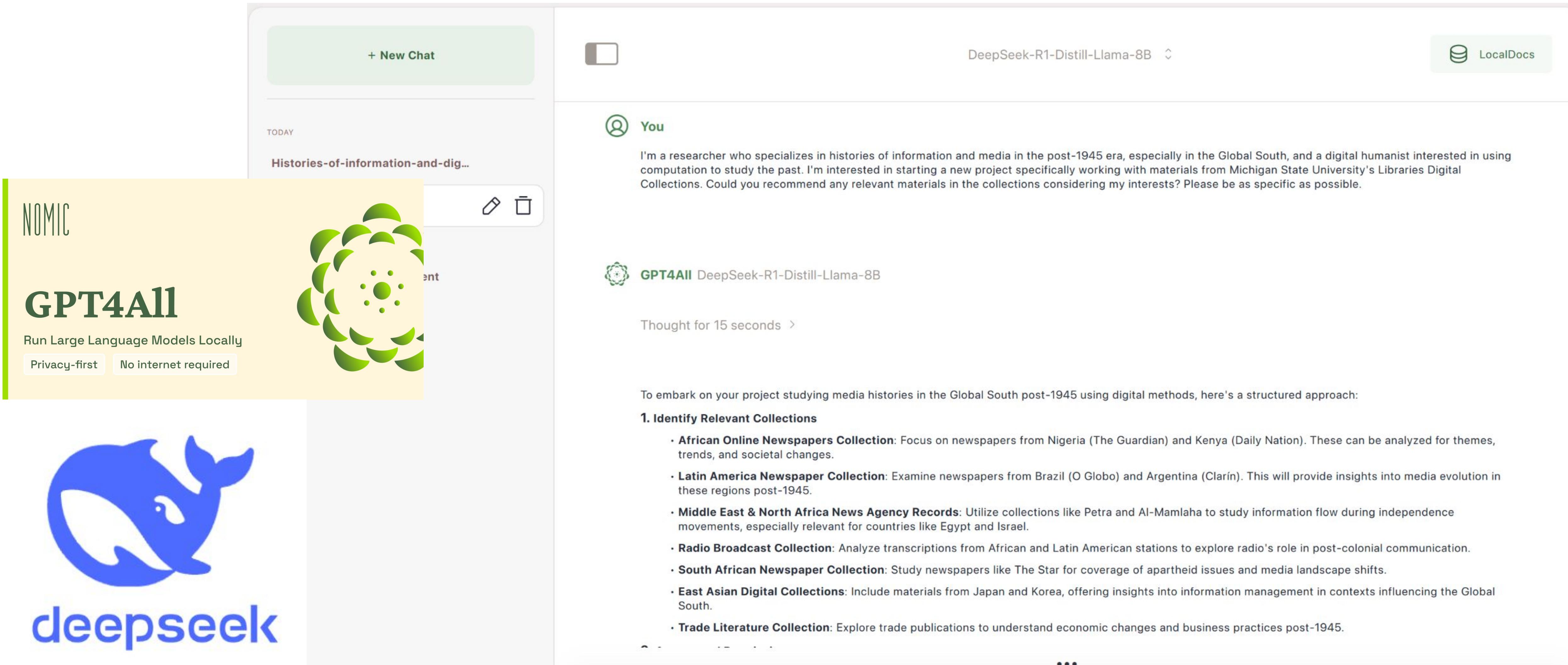
Pretrained

Trained on a simple task, but can be applied to many others!

Transformer

The ML model that guesses the word.

Initial Results With GPT4ALL & DeepSeek



The screenshot shows a split-screen interface. On the left, the GPT4All application is displayed. It features a green header bar with the text "NOMIC" and "GPT4All". Below this is a yellow bar with the text "Run Large Language Models Locally", "Privacy-first", and "No internet required". The main area contains a large green logo of a brain-like structure composed of many smaller circles. To the right of the logo, there is a text input field with the placeholder "Histories-of-information-and-dig...". On the far right of the GPT4All interface, there are two small buttons: a pencil icon and a trash bin icon.

The right side of the screenshot shows a chat interface. At the top, it says "DeepSeek-R1-Distill-Llama-8B". In the middle, there is a message from "You" which reads: "I'm a researcher who specializes in histories of information and media in the post-1945 era, especially in the Global South, and a digital humanist interested in using computation to study the past. I'm interested in starting a new project specifically working with materials from Michigan State University's Libraries Digital Collections. Could you recommend any relevant materials in the collections considering my interests? Please be as specific as possible." Below this, there is a message from "GPT4ALL DeepSeek-R1-Distill-Llama-8B" which says "Thought for 15 seconds >". Further down, there is a structured approach for the research project:

To embark on your project studying media histories in the Global South post-1945 using digital methods, here's a structured approach:

1. Identify Relevant Collections

- **African Online Newspapers Collection:** Focus on newspapers from Nigeria (The Guardian) and Kenya (Daily Nation). These can be analyzed for themes, trends, and societal changes.
- **Latin America Newspaper Collection:** Examine newspapers from Brazil (O Globo) and Argentina (Clarín). This will provide insights into media evolution in these regions post-1945.
- **Middle East & North Africa News Agency Records:** Utilize collections like Petra and Al-Mamlaha to study information flow during independence movements, especially relevant for countries like Egypt and Israel.
- **Radio Broadcast Collection:** Analyze transcriptions from African and Latin American stations to explore radio's role in post-colonial communication.
- **South African Newspaper Collection:** Study newspapers like The Star for coverage of apartheid issues and media landscape shifts.
- **East Asian Digital Collections:** Include materials from Japan and Korea, offering insights into information management in contexts influencing the Global South.
- **Trade Literature Collection:** Explore trade publications to understand economic changes and business practices post-1945.

At the bottom right of the chat interface, there is a small "..." button.

Initial Results With Perplexity AI

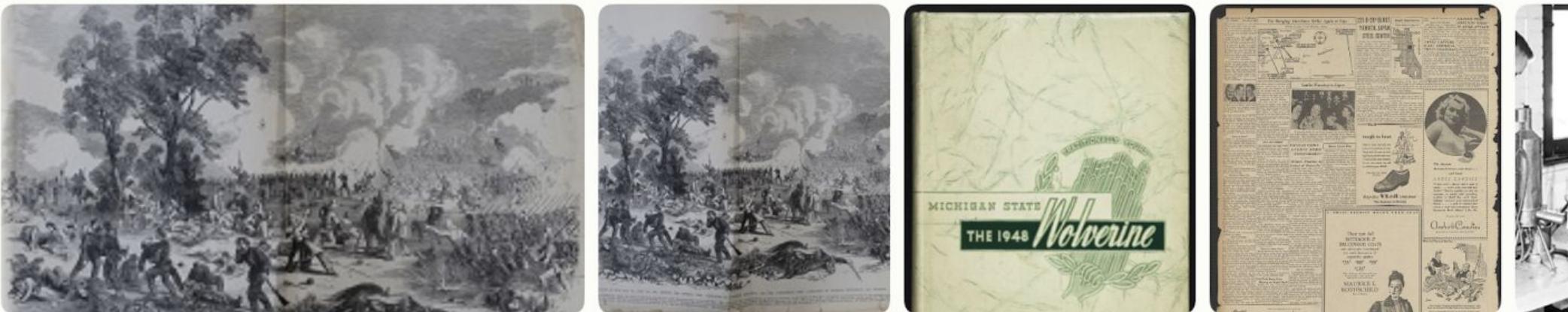


I'm a researcher who specializes in histories of information and media in the post-1945 era, especially in the Global South, and a digital humanist interested in using computation to study the past. I'm interested in starting a new project specifically working with materials from Michigan State University's Libraries Digital Collections. Could you recommend any relevant materials in the collections considering my interests? Please be as specific as possible

Sources



Answer



Michigan State University's Libraries Digital Collections offer a wealth of materials that align with your interests in the histories of information, media, and the Global South, as well as your focus on computational methods for research and study. Here are some specific collections that may be particularly relevant:

1. MSU Vietnam Group Archive

- **Content:** Documents, maps, and images from 1955–1962 related to Michigan State University's involvement in the Vietnam War.
- **Relevance:** This collection provides insight into post-1945 development initiatives and U.S. influence in Southeast Asia, making it an excellent resource for studying media representations and knowledge production.

1 3

2. Voices of MSU

- **Content:** Oral histories featuring interviews with minority students, faculty, and staff at MSU.
- **Relevance:** These interviews could be valuable for exploring narratives of marginalized groups in the Global South.

Ask follow up

Recommendations for Digital Humanities Approaches:

- Employ **text mining** or **topic modeling** to analyze large textual datasets like oral histories or yearbooks.
- Use **network analysis** to map relationships between individuals or institutions in the Vietnam Group Archive.
- Apply **visual recognition algorithms** to study imagery trends in yearbooks or photographs.

These collections provide a robust foundation for exploring how information and media have been produced, preserved, and represented in different historical contexts.

↑ ↵ ⏪ ...

Related

- Q What specific collections in the University Archives & Historical Collections are most relevant to my research on the Global South
- Q Are there any digitized materials from the 1950s and 1960s that could provide insights into post-1945 history in the Global South
- Q Can I access interviews with minority students, faculty, and staff from Michigan State University's archives
- Q What computational methods can I use to analyze the digitized materials from MSU's Libraries Digital Collections
- Q Are there any historical records or documents related to academic programs at MSU that could be useful for my research

Ask follow up

Pro

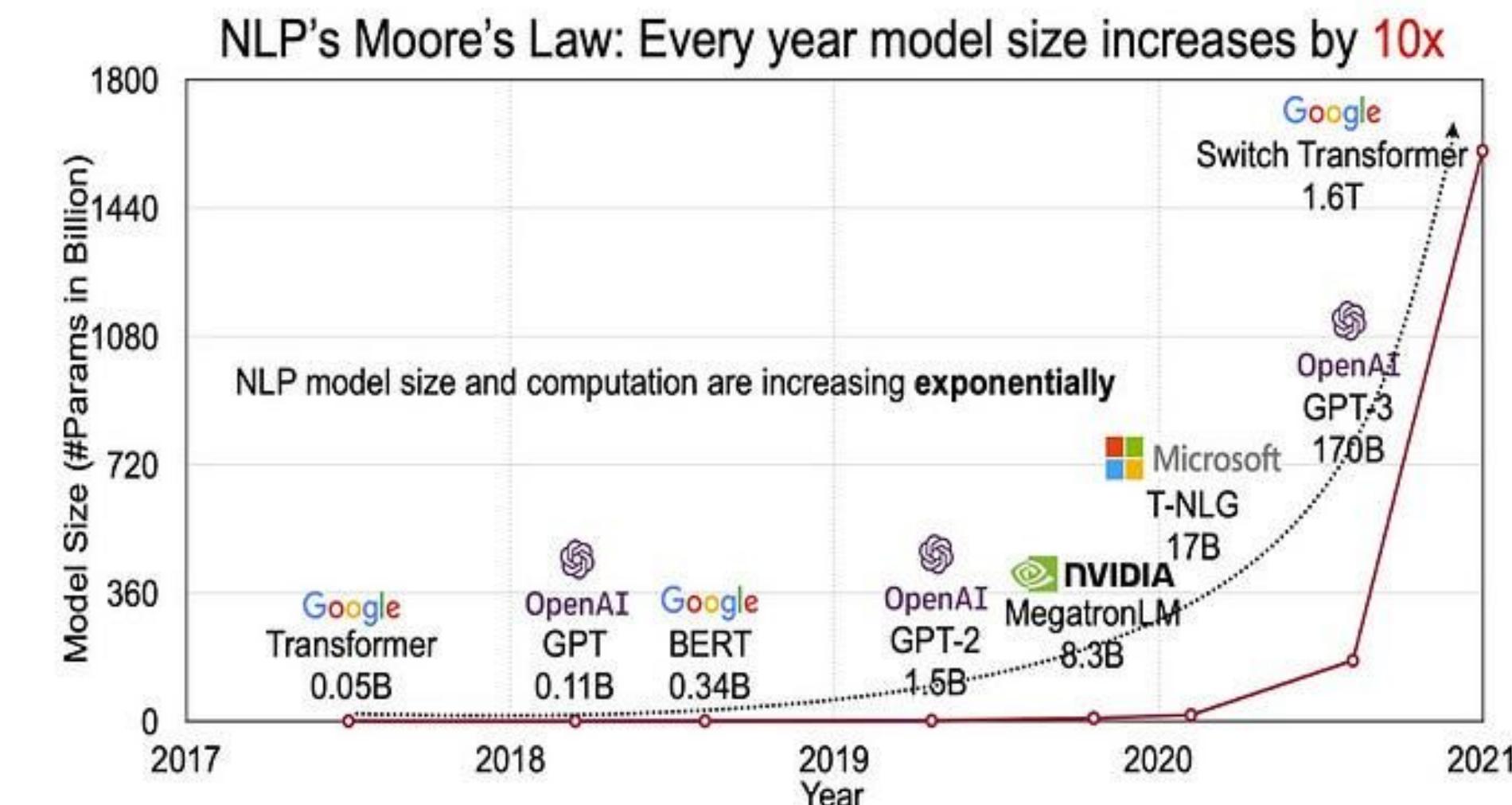
Distinguishing Between Chatbots: Web Search & Model Parameters

● Increasingly Search-Enabled

- Early chatbots were limited to pre-trained knowledge, but many now incorporate web search, allowing for more real-time and dynamic responses.
- Not all models have equal access—some prioritize proprietary databases, while others browse the open web.

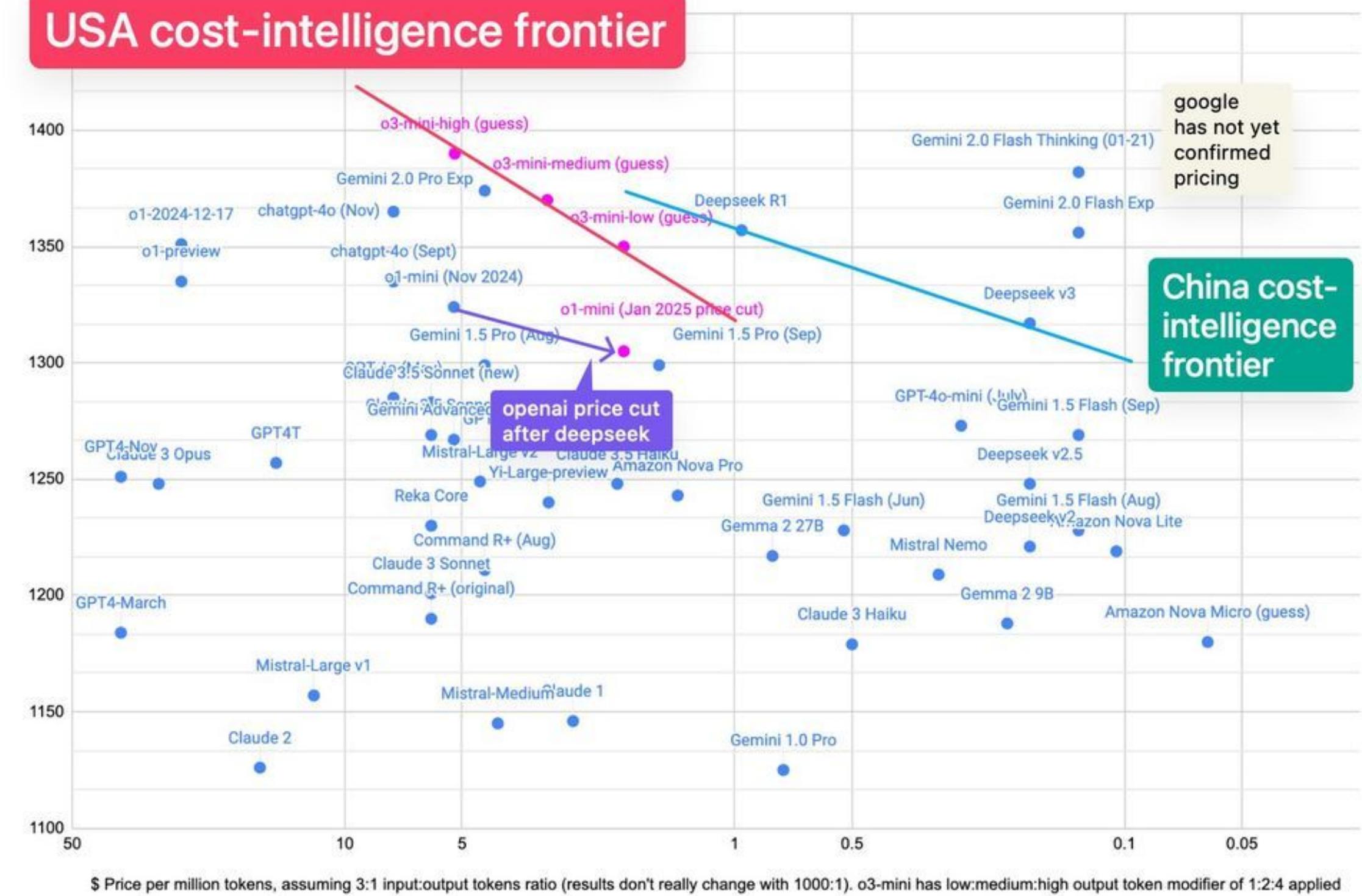
● LLMs Follow Moore's Law

- The cost of training models continues to decrease as hardware efficiency and algorithmic improvements advance.
- Model size and capability are growing exponentially, but efficiency gains mean some models are now small enough to run locally on phones or personal computers.



Plot of model pricing vs LMSys Elo (Jan 2025) - more on @latentspacepod

USA cost-intelligence frontier



Evaluating Initial Prompt

Chatbot	Model	Found Relevant Materials?	Suggested Helpful Methods?	Costs	Data Privacy	Link to Log
Claude Anthropic	Sonnet 3.5	Not really. Very generic answer.	Minimal	Free Tier	Data is supposed to be private unless opted in to train models. More info here A I would like to input sensitive data into Free Claude.ai or ...	https://shareclaude.pages.dev/c/f7sigzkfck16w9meitbesjvs
Microsoft Co-Pilot	Microsoft Prometheus model, which combines the Bing search index with OpenAI's GPT-4 models.	Not really. Very generic answer.	None	Institutional License via UIUC	Data is encrypted and web search queries are removed after 24 hours.	
GPT4ALL	DeepSeek	Not really. Very generic answer.	Minimal	Free	Complete private and running locally	
Perplexity AI	Default Model	Found some relevant materials and included links.	Suggested some methods, but primarily resources.	Free Tier	Defaults to using user queries to train models, but can opt out in settings.	
Google Gemini	2.0 Flash Thinking Experimental	Found some relevant materials and included links.	None	Free Tier	Claims that user queries are not used to train models. More info here  Google Cloud How Gemini for Google Cloud uses your data	 I'm interested in starting a new project specific...
OpenAI ChatGPT	4o	Found some relevant materials and included links.	Mentioned resources but no methods.	Plus Tier (\$20 USD/Month)	Claims that users can opt-in to having their data used to train models. More info here  Data Controls FAQ OpenAI Help Center	 ChatGPT ChatGPT - MSU Digital Collections Research
OpenAI ChatGPT	o1 pro	Found some relevant materials and included links. More detailed than the 4o model.	Mentioned some methods and resources.	Pro Tier (\$200 USD/Month)	Claims that users can opt-in to having their data used to train models. More info here  Data Controls FAQ OpenAI Help Center	<i>My assessment: essentially provided generic search engine results.</i>

Evaluating Revised Prompt

Chatbot	Model	Found Relevant Materials?	Suggested Helpful Methods?	Costs	Data Privacy	Link to Log
Claude Anthropic	Sonnet 3.5	Yes, but no links.	Suggested using OCR with some text analysis, so somewhat useful.	Free Tier	Data is supposed to be private unless opted in to train models. More info here A I would like to input sensitive data into Free Claude.ai or ...	https://shareclaude.pages.dev/c/h6ypejdxqwrypm4ptkpracev
Microsoft Co-Pilot	Microsoft Prometheus model, which combines the Bing search index with OpenAI's GPT-4 models.	Yes with links.	None.	Institutional License via UIUC	Data is encrypted and web search queries are removed after 24 hours.	
GPT4ALL	DeepSeek	Somewhat, but still fairly generic and no links.	Suggested both specific methods and Python libraries.	Free	Complete private and running locally	
Perplexity AI	Default Model	Technically found some potentially relevant collections, but not much better than initial prompt.	Suggested some methods but not very detailed.	Free Tier	Defaults to using user queries to train models, but can opt out in settings.	
Google Gemini	2.0 Flash Thinking Experimental	Yes with links.	Suggested relevant methods.	Free Tier	Claims that user queries are not used to train models. More info here  Google Cloud How Gemini for Google Cloud uses your data	 I'm a researcher who specializes in histories of...
OpenAI ChatGPT	4o	Yes with links.	Suggested some resources.	Plus Tier (\$20 USD/Month)	Claims that users can opt-in to having their data used to train models. More info here  Data Controls FAQ OpenAI Help	 ChatGPT ChatGPT - MSU Digital Collections Research
OpenAI ChatGPT	o1 pro	Most detailed response with links.	Suggested methods for each collection.	Pro Tier (\$200 USD/Month)	<p><i>Revised prompt had much more accurate results. Still essentially doing search but with a lot more contextual and specific assistance.</i></p>	

Center

Success? Accuracy vs. Automation

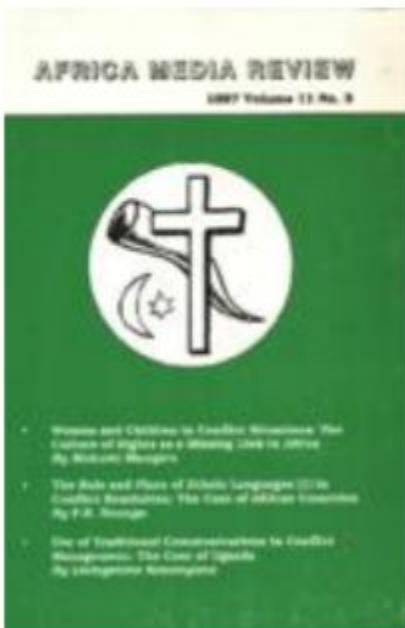
MSU Libraries | Digital Repository

[Home](#) [About](#) [Collections](#)



[Home](#) / [MSU Libraries Digital Collections](#)

Africana Studies Collection

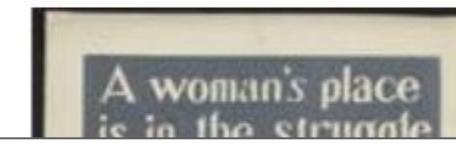


African e-Journals
Project



Africana Libraries
Newsletter

[View 3 collections](#)



2. African e-Journals Project

- What It Is:** A collaborative digitization project providing full-text access to selected academic and scholarly journals published in Africa. While many issues span the 1960s through the 2000s, coverage varies by title.
- Types of Sources:** Peer-reviewed journals, research bulletins, occasional papers in fields such as history, political science, communications, sociology, and literature—many of them difficult to find elsewhere.
- Why It's Relevant:**
 - Excellent window into how knowledge and research circulated within Africa from the early postcolonial period onward.
 - Journals sometimes contain book reviews, commentary on media and information policies, and debates on communication infrastructures in newly independent states.
 - Ideal for large-scale text mining to study discourse trends (e.g., how "information policy" or "mass media" is discussed over time).

Where to Start: [African e-Journals Project](#)

Part 3

Collecting & Curating Research Materials With AI

*How can AI help us discover relevant materials and situate our questions?
How can we evaluate and trust the results from these models?*

Initial Results With Google's Gemini

Gemini ▾
2.0 Flash

- Gemini
- 2.0 Flash Get everyday help New
- 2.0 Flash Thinking Experimental Best for multi-step reasoning New
- 2.0 Flash Thinking Experimental with apps Reasoning across YouTube, Maps & Search New

Gemini Advanced Upgrade

I'm a researcher who specializes in histories of information and media in the post-1945 era, particularly in the Global South, and a digital humanist

Show thinking ▾

- ⓘ You're using 2.0 Flash Thinking Experimental with reasoning across best with YouTube, Maps, and Search. Some Gemini features aren't available yet.

Michigan State University (MSU) Libraries offers a wealth of digital collections relevant to your research interests. Based on your specialization in histories of information and media in the post-1945 era, particularly in the Global South, and your interest in digital humanities, here are some specific collections and materials within MSU Libraries Digital Collections that you might find valuable:

Digital Repository

The MSU Libraries Digital Repository is a great starting point. It hosts a variety of collections that could align with your interests. You can access it here: [MSU Libraries Digital Collections - Digital Repository](#). Some potentially relevant collections within the Digital Repository include:

Ask Gemini



Gemini can make mistakes, so double-check it

Show thinking ^

The user is a researcher interested in histories of information and media in the post-1945 Global South and digital humanities, and wants to know about relevant materials in Michigan State University's Libraries Digital Collections. I need to search Michigan State University's Libraries Digital Collections and identify collections relevant to the user's interests. I will use Google Search to find Michigan State University's Libraries Digital Collections and then search within the collections for relevant materials. I should focus on collections related to post-1945, Global South, information, media, and potentially digital humanities.

First, I will search for "Michigan State University Libraries Digital Collections" to find the link to the collections. Then, I will explore the collections and identify relevant ones based on the user's interests. I will try to be as specific as possible in my recommendations, naming specific collections and potentially even sub-collections or types of materials within those collections.

Search query: "Michigan State University Libraries Digital Collections"

The search results show two main links for MSU Libraries Digital Collections:

1. <https://www.library.msstate.edu/digitalcollections> (MSU Libraries Collections)
2. <https://lib.msu.edu/dmc/collections> (Michigan State University Gerald M. Kline Digital & Multimedia Center)
3. <https://d.lib.msu.edu/msul> (MSU Libraries Digital Collection - Michigan State University)

It seems there might be confusion as the first link is for Mississippi State University. The user is interested in Michigan State University. The second and third links seem to be for Michigan State University. I should focus on the second and third links.

Let's examine the second link <https://lib.msu.edu/dmc/collections> first as it is directly named "Browse Digital Collections - MSU Libraries - Michigan State University". This page lists collections such as Africana, American Radicalism, Comic Art, History and Social Sciences, etc. "Africana" and "History and Social Sciences" could be relevant to the user's interest in the Global South and post-1945 era. "American Radicalism" might be relevant to media and information history.

Let's also check the third link <https://d.lib.msu.edu/msul> which is "MSU Libraries Digital Collections - Digital Repository - Michigan State University". This page mentions collections like "Electronic Theses & Dissertations", "Michigan Farming Historical Newspapers", "Radicalism Posters Collection", "G. Robert Vincent Voice Library Collection". "Radicalism Posters Collection" and "G. Robert Vincent Voice Library Collection" might be relevant.

I need to explore these collections in more detail to see if they are relevant to the user's interests in histories of information and media in the post-1945 era, especially in the Global South, and digital humanities.

I will start by Browse the "Browse Digital Collections - MSU Libraries" page (<https://lib.msu.edu/dmc/collections>) and look for collections that seem relevant. I will then browse the "Digital Repository" page (<https://d.lib.msu.edu/msul>) as well.

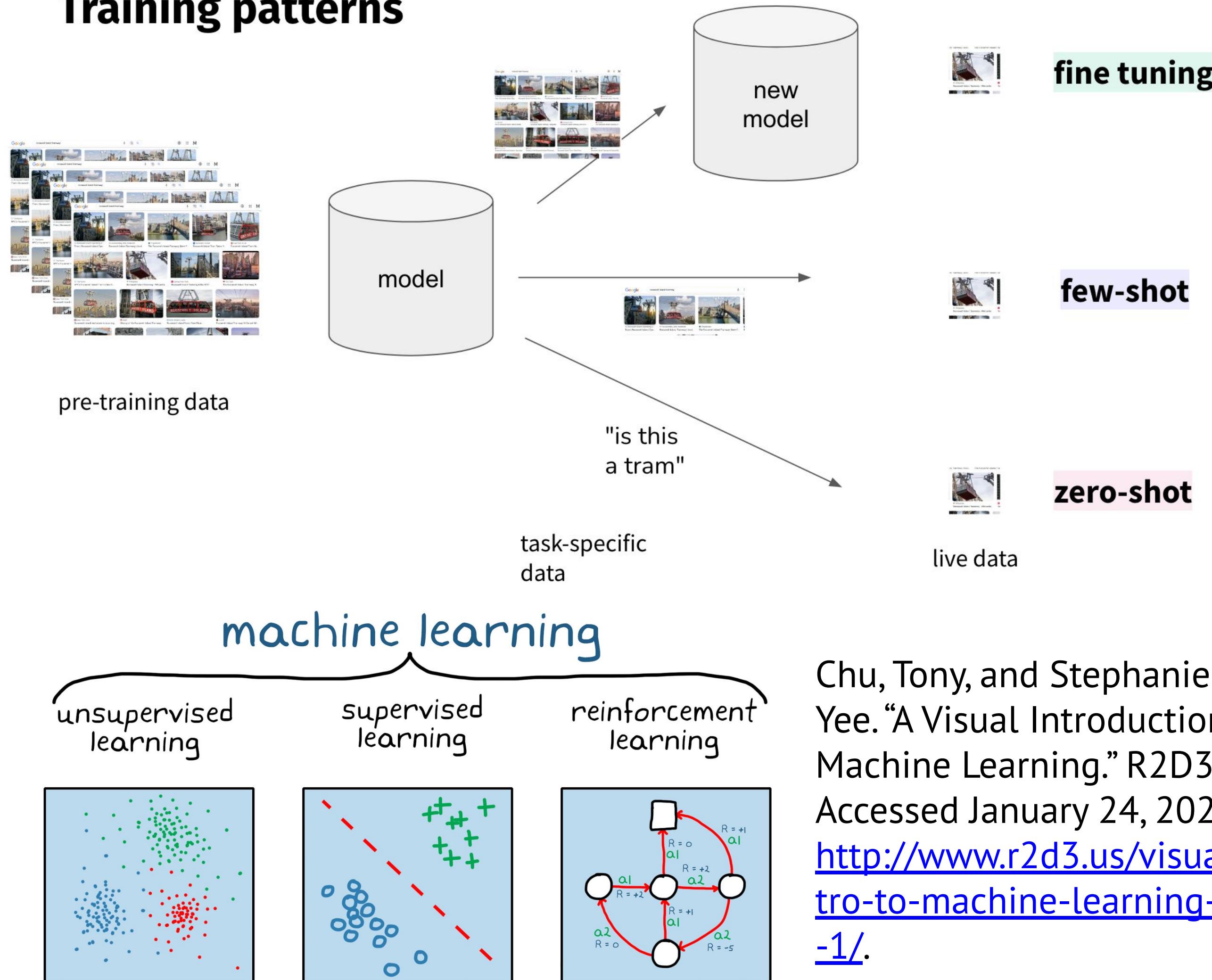
I will start with the "Browse Digital Collections - MSU Libraries" page and browse the "Africana" collection to see if it has materials related to post-1945 and media/information in the Global South.

Browse query for <https://lib.msu.edu/dmc/collections> : "Africana collection description"

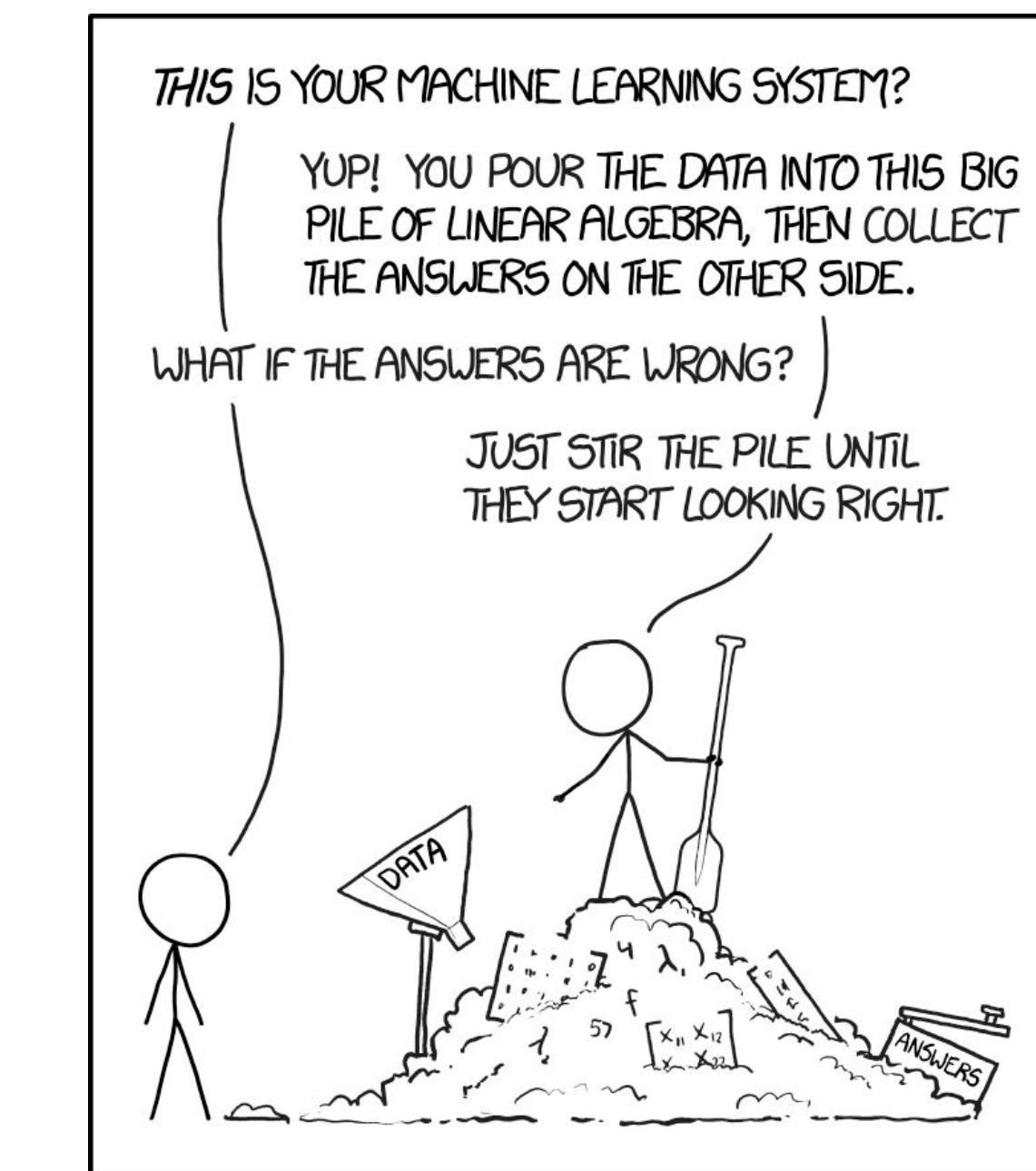
See chat log here
I'm a researcher
who specializes in
histories of...

Core Concepts With AI: Learning Methods

Training patterns



Chu, Tony, and Stephanie Yee. "A Visual Introduction to Machine Learning." R2D3. Accessed January 24, 2025. <http://www.r2d3.us/visual-intro-to-machine-learning-part-1/>.



The terms “data science,” “neural networks,” “artificial intelligence,” and “machine learning” are often deployed interchangeably, but all, especially “AI,” tend to obfuscate what algorithms are doing today. The building hype over artificial intelligence in the last decade has been largely in a specific area that I have found most usefully described as “representation learning.”

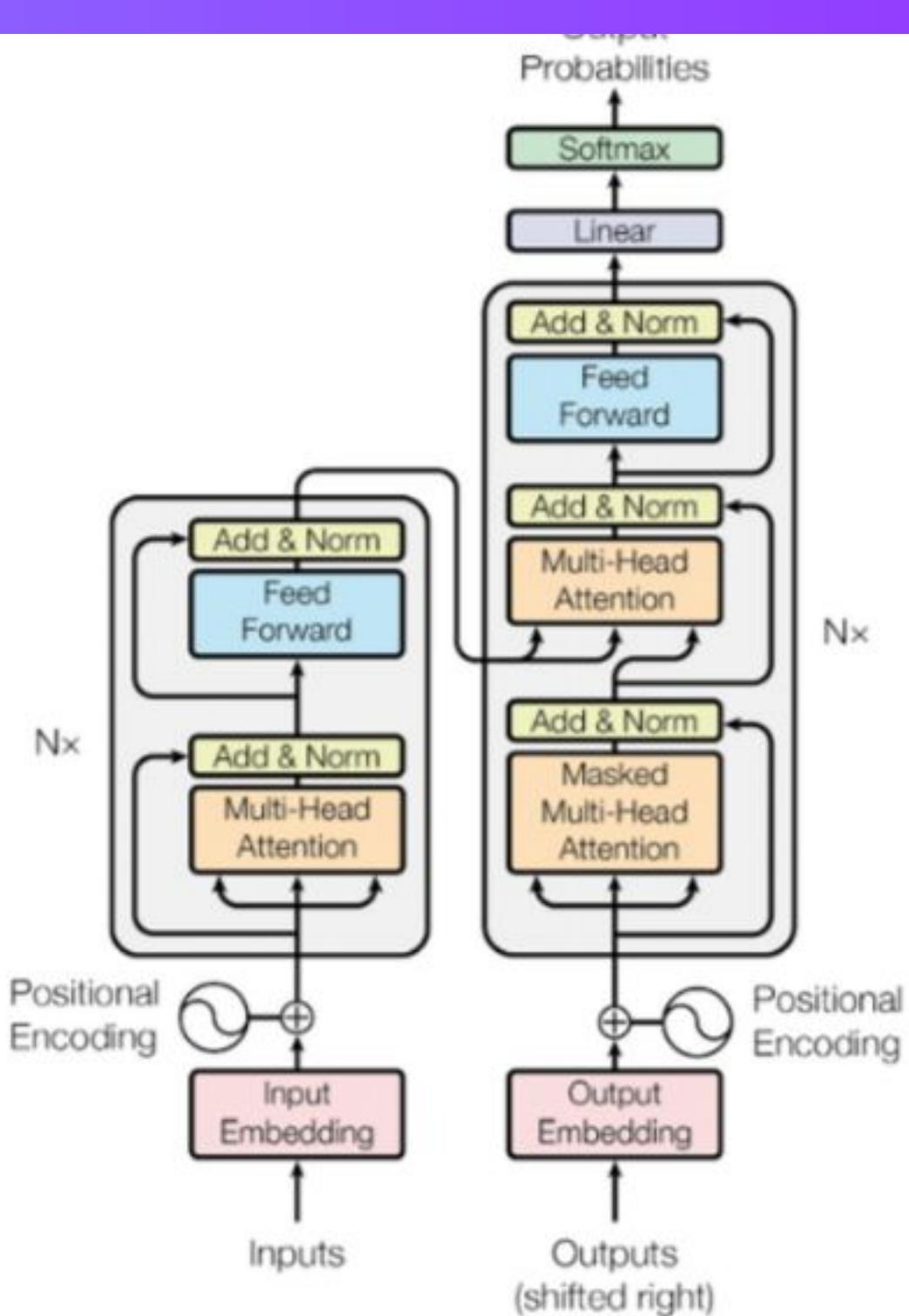
Benjamin Schmidt, “Representational Learning” in “Artificial Intelligence and the Practice of History: A Forum.” *The American Historical Review* 128, no. 3 (September 1, 2023): 1345–49. <https://doi.org/10.1093/ahr/rhd362>.

Core Concepts with AI: Transformers & Foundation Models

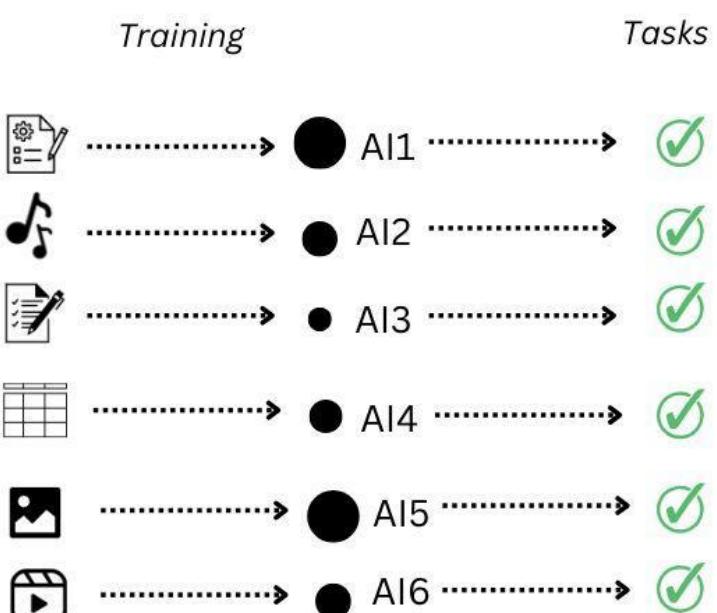
Through learning, artificial intelligence systems are doing much more than improving an operation. Like historians, they gather historical data and classify, analyze, interpret, and store it for future use. This intricate series of activities cannot occur without historical awareness that draws upon memory, critical analysis, data contextualization, and a theory of causality. I call this phenomenon the **AI-as-Historian**.

Transformer

Attention Is All You Need

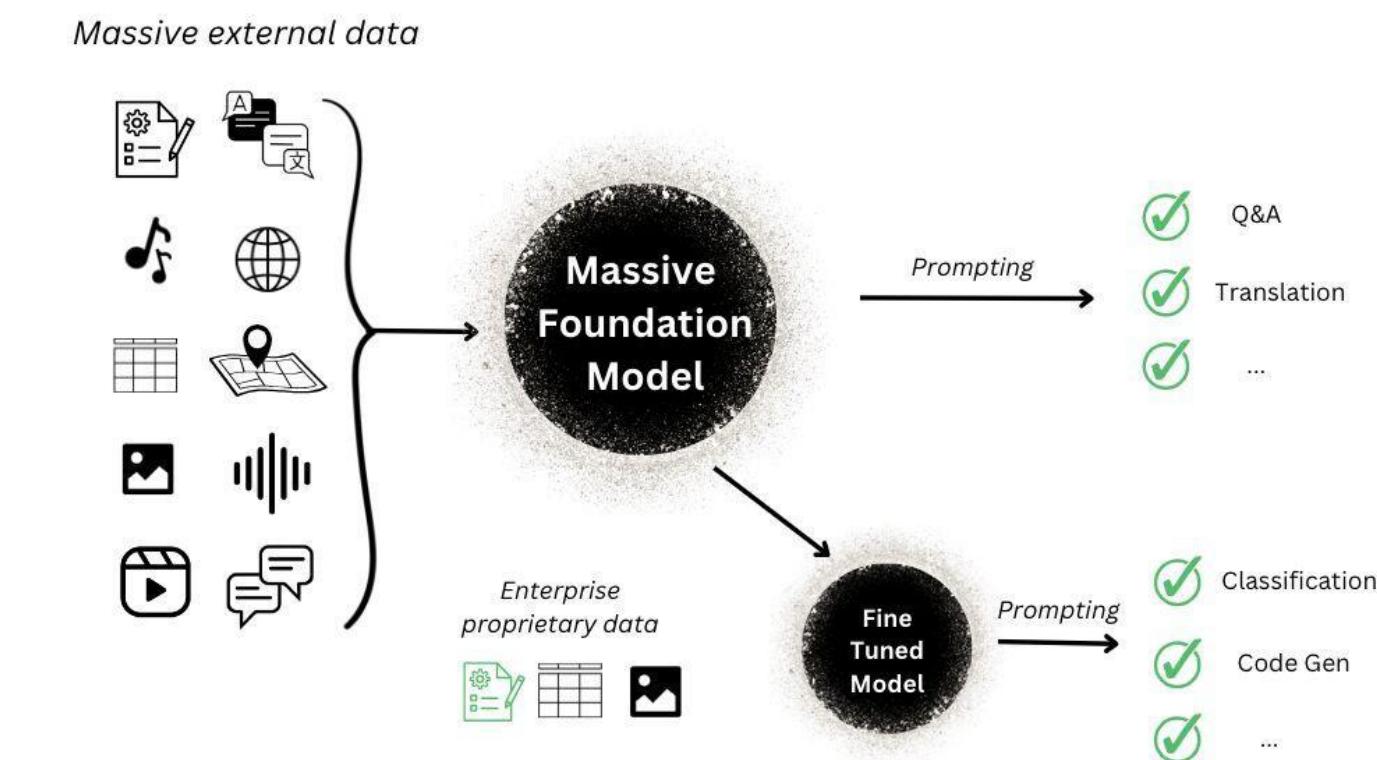


Traditional ML



- Individual siloed models
- Require task-specific training
- Lots of human supervised training

Foundation Models



- Massive multi-tasking model
- Adaptable with little or no training
- Pre-trained unsupervised learning

Joshua Sternfeld, "AI-as-Historian" in "Artificial Intelligence and the Practice of History: A Forum." *The American Historical Review* 128, no. 3 (September 1, 2023): 1345–49.
<https://doi.org/10.1093/ahr/rhd362>.

Core Concepts With AI: Deep Research & Reinforcement Learning



An unexpected
RL Renaissance

Nathan Lambert
Allen Institute for AI (& Interconnects.ai)
Minds & Machines, Seattle WA
12 February 2025

Lambert | RL Renaissance 1

The screenshot shows the OpenAI website with the Deep Research product page. The title is "Introducing deep research". Below it is a description: "An agent that uses reasoning to synthesize large amounts of online information and complete multi-step research tasks for you. Available to Pro users today, Plus and Team next." A "Try on ChatGPT" button is visible. The date is February 2, 2025.

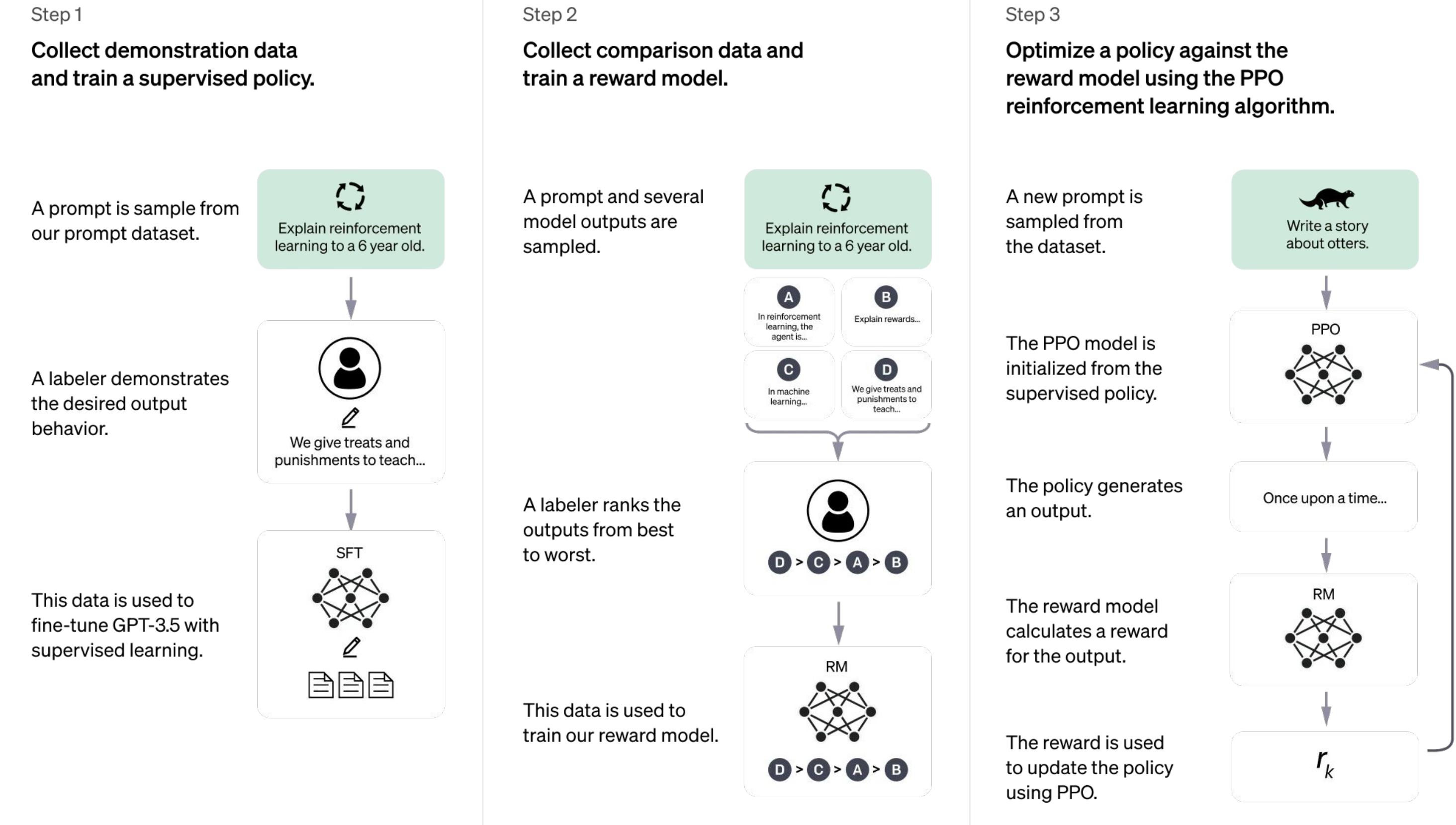
The End of Search, The Beginning of Research

The first narrow agents are here

FEB 03, 2025

768 55 70

A hint to the future arrived quietly over the weekend. For a long time, I've been discussing two parallel revolutions in AI: the rise of autonomous agents and the emergence of powerful Reasoners since OpenAI's o1 was launched. These two threads have finally converged into something really impressive - AI systems that can conduct research with the depth and nuance of human experts, but at machine speed. OpenAI's Deep Research demonstrates this convergence and gives us a sense of what the future might be. But to understand why this matters, we need to start with the building blocks: Reasoners and agents.



Reasoners



GPT4ALL DeepSeek-R1-Distill-Llama-8B

Thought for 13 seconds ▾

ChatGPT 01 pro >

Research Materials on Global South History

Reasoned for 5m 24s

Below are several Michigan State University Libraries Digital Collections that focus on post-1945 histories of information and media—especially in the Global South. I've highlighted specific strengths, date ranges, and types of media and information flows.

1. African Activist Archive

- **What It Is:** A digital repository of more than 10,000 items documenting African political movements, media coverage of democracy, roughly from the 1950s to the 1990s. Although it primarily focuses on African political movements, it also includes reference to African political movements, media coverage of the Cold War, and other global issues.
- **Types of Sources:** Newsletters, leaflets, pamphlets, buttons, posters, photographs, audio recordings, and video recordings from various organizations.

I'm a researcher focusing on post-1945 histories of information and media, and perspectives from the Global South. I'm exploring Michigan State University's digital collections, including African studies, diaspora, Afrobarometer data, and the African Activist Archive, which might shed light on activism.

I'm gathering details about Michigan State University's digital collections, like the African Activist Archive, G. Robert Vincent Voice Library, and the Vietnam Group Archive, which could boost the digital humanities perspective.

I'm exploring digitized newspapers, African e-journals, and the Africana Pamphlet Collection to examine media from the post-1945 era in the global south through a digital humanist lens.

Focusing on MSU's digital collections that align with post-1945 global south media studies. Highlighting the African Activist Archive and Africana Collection.

Okay, so I'm a historian of information and digital humanist looking to work with Michigan State University's digitized collections. Hmm, where do I start? First off, I know that MSU has some pretty cool resources, but I need to figure out which ones would be most relevant for my fields.

I remember hearing about the Digital Library for Earth Sciences at MSU. That sounds like it could have a lot of historical data on things like climate change over time. Maybe they have old maps or weather records? That would definitely tie into understanding past environmental changes, which is important for both information history and digital humanities projects.

Then there's the African American History Collection. I think that includes documents, photographs, and oral histories from various periods. This could be useful if I'm looking at how communities of color have interacted with technology or information over time. It might also provide insights into social movements related to civil rights, which is a big part of digital humanist studies.

The Special Collections in the MSU Libraries probably hold rare books and manuscripts. That's perfect for someone interested in the history of information because it includes first editions and unique materials that show how knowledge was disseminated before the digital age. Maybe there are letters from famous scientists or historical figures, which would be great sources for research.

The Michigan State Normal School Collection is another one I've heard about. It seems to focus on education and has a lot of 19th-century documents. If I'm looking at how information was taught back then, this could provide valuable materials. Maybe there are teaching manuals or student notebooks that show the evolution of educational practices.

The United States Census Bureau's Historical Statistics might be useful for demographic studies. It includes population data from past censuses, which can help in analyzing social and economic trends over time. This is relevant to both information history (how data was collected) and digital humanities projects on societal changes.

Moving Image Collection sounds interesting too. Visual media like photographs and films could offer insights into cultural shifts or technological advancements. For a digital humanist, looking at how visual content has evolved might provide new angles for analysis in their work.

The MSU Archives holds university records, which can be goldmines for understanding the history of higher education. If I'm interested in how universities have adapted to changes over time, especially regarding technology adoption or curriculum development, this would be a great resource.

I also need to check out the Michigan State University Digital Collections website. They probably have an online portal where these collections are accessible. It might require some digging to find specific materials, but that's part of the research process.

Wait, I should make sure if there are any restrictions or access requirements for these collections. Some might require permissions or be limited in how they can be used. Also, considering the scope of each collection, I need to prioritize which ones align most with my current projects and interests.

Prompting With Deep Research

Generative History

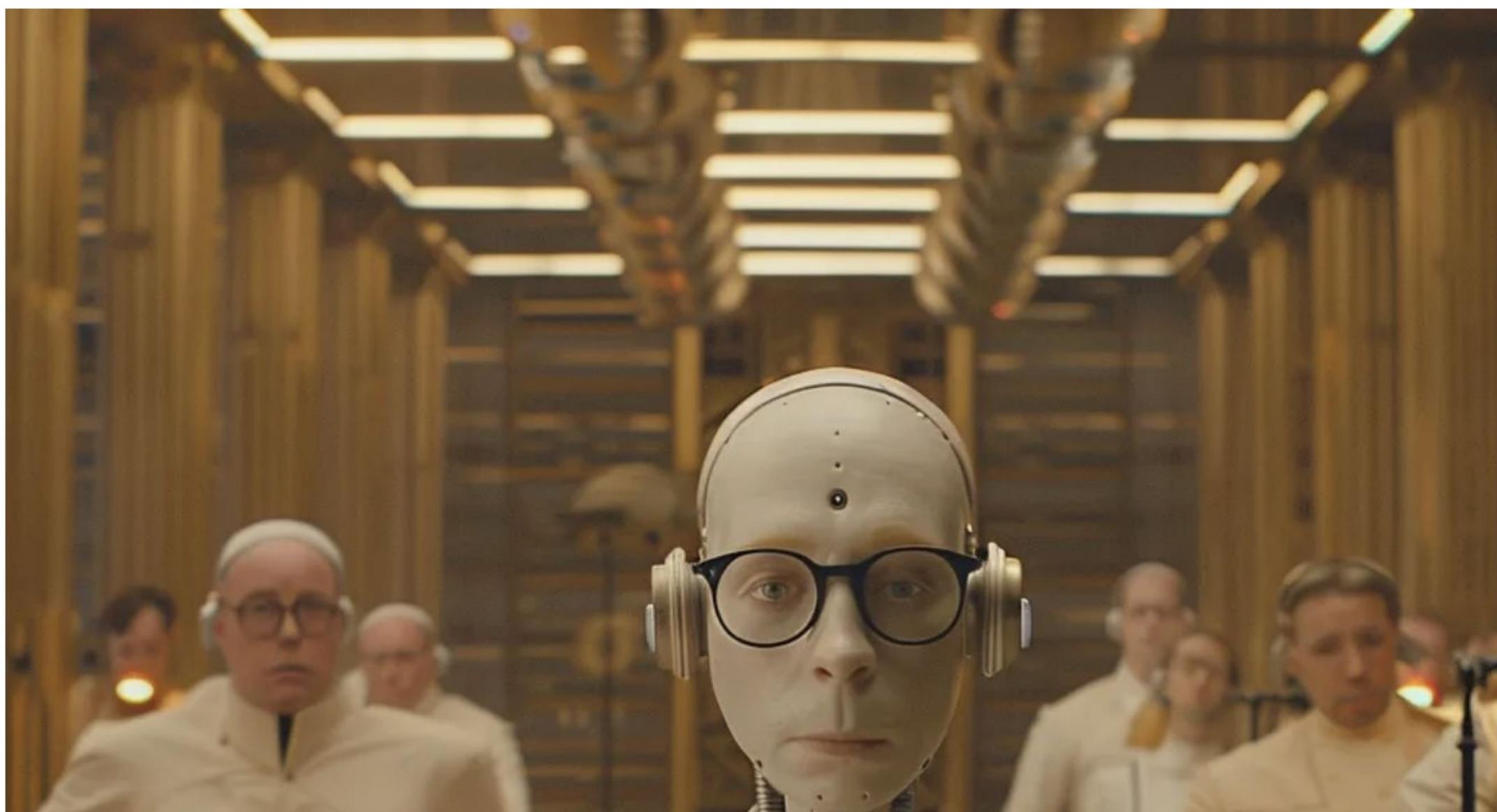
Is this the Last Generation of Historians?

OpenAI's Deep Research is a strange and shockingly powerful agentic AI research assistant that offers a clear glimpse of the future. But where do we fit in?



MARK HUMPHRIES
FEB 07, 2025

36 29 7



Ethan Mollick
@emollick.bsky.social

There are no rigorous studies yet on how to best prompt reasoning models like o1.

They are not chat models, they are designed so that you give them a question and you get an answer. You want to give them context, clarity, etc. but we don't know what the optimal way to do that is yet.

January 19, 2025 at 12:35 PM · Everybody can reply

Historiographical Analysis of Fur Trade Families (1763–1821)

Introduction

The history of fur trade families in North America (1763–1821) has given rise to two distinct historiographical traditions, shaped in large part by Canadian and American national narratives.

Canadian scholars have long treated the fur trade as foundational to the country's development, whereas American historians traditionally viewed it as a peripheral prelude to agrarian expansion

DEGRUYTER.COM LIBERATINGNARRATIVES.COM Over time, however, these divergent approaches have converged on key themes—such as the role of Indigenous women, the emergence of Métis (mixed-descent) communities, and the importance of kinship networks—thanks to a shared turn toward social history and ethnohistory. This essay surveys major works in the field, comparing Canadian and American historiographical approaches to fur trade families. It highlights central debates and methodological differences, and explains how national origin myths initially drove divergent interpretations even as recent scholarship bridges those “glass curtains” of the 49th parallel DEGRUYTER.COM Key scholars in both traditions—Harold Innis, Sylvia Van Kirk, Jen S.H. Brown in Canada; Hiram Chittenden,

Message ChatGPT

+

Search

83



...

Prompting With Deep Research

Write a historiographical essay contextualizing the journal, *Africa Media Review*, specifically within histories of information and decolonization. I also want you to use digital humanities methods where appropriate. Be sure to only use peer-reviewed academic sources and be sure to refer to key historians by name. Aim to be comprehensive.

Write a historiographical essay comparing and contrasting Canadian and American approaches to the history of fur trade families in North America. Be sure to only use peer-reviewed academic secondary sources and be sure to refer to key historians by name. Aim to be comprehensive.

BLANK PAGE APPROACH

Initial Results With OpenAI's ChatGPT Deep Research o1 pro

ChatGPT 4o mini ▾

This is a copy of a conversation between ChatGPT & Anonymous.

[Report content](#)

Write a historiographical essay contextualizing the journal, Africa Media Review, specifically within histories of information and decolonization. I also want you to use digital humanities methods where appropriate. Be sure to only use peer-reviewed academic sources and be sure to refer to key historians by name. Aim to be comprehensive.

Africa Media Review and Decolonization

Reasoned for 10m 42s

A Historiographical Essay on *Africa Media Review* within the Histories of Information and Decolonization

The journal *Africa Media Review*, founded in 1986 by the African Council for Communication Education (ACCE), occupies a critical yet often underexplored position within the broader histories of African knowledge production, decolonization, and the transformation of global information flows. While much attention in historiography has been devoted to the political, economic, and social dimensions of Africa's decolonization, *Africa Media Review* reminds us that the decolonization of information—encompassing media, communication studies, and cultural representation—remains a vital dimension in understanding Africa's postcolonial trajectories.

Message ChatGPT



Details

Tracing Africa Media Review's origins

I'm noting the African Council for Communication Education (ACCE) launched Africa Media Review in the 1980s or 1990s, aligning with African media studies, Pan-African efforts, UNESCO's NWICO, and ICT4D.

Piecing together media evolution

I'm leveraging historical figures like Daniel Lerner and Walter Rodney to trace African media's scholarly growth. Also, I'm integrating digital humanities methods for analysis.

▼ First Attempt

A Historiographical Essay on *Africa Media Review* within the Histories of Information and Decolonization

The journal *Africa Media Review*, founded in 1986 by the African Council for Communication Education (ACCE), occupies a critical yet often underexplored position within the broader history of African knowledge production, decolonization, and the transformation of global information flows. While much attention in historiography has been devoted to the political, economic, and social dimensions of Africa's decolonization, *Africa Media Review* reminds us that the decolonization of information—encompassing media, communication studies, and cultural representation—remains a vital dimension in understanding Africa's postcolonial trajectories. This essay seeks to contextualize *Africa Media Review* within key debates on decolonizing knowledge in African history, situate its contributions in relation to major historians' interpretations, and outline how digital humanities methods can further illuminate the journal's role in reshaping narratives about African media and communication.

1. Decolonization and the Politics of Knowledge Production

1.1. Early Historiographical Perspectives

For much of the twentieth century, narratives surrounding Africa's colonial and postcolonial histories were dominated by perspectives emanating from Western academic institutions. As the historians Jean Suret-Canale (1964) and Roland Oliver (1961) made clear in their foundational studies of colonial Africa, the colonial archive itself was shaped by European priorities and biases. Even with the rise of African nationalist historiography in the mid-twentieth century, scholars like Kenneth Dike (1956) and Bethwell A. Ogot (1963) insisted on the need

See chat log here
<https://chatgpt.com/share/67b7b95e-d504-8006-99b6-434ec9784a9b>

▼ Second Attempt

A Historiographical Essay on *Africa Media Review* within the Histories of Information and Decolonization

Since its founding in 1986 under the auspices of the African Council for Communication Education (ACCE), *Africa Media Review* has provided a forum for examining African media and communication through African-centered lenses. Despite the journal's somewhat modest profile in broader academic discourse, its significance for understanding decolonization—particularly the decolonization of information—cannot be overstated. This essay offers a historiographical overview of *Africa Media Review*, embedding it within critical debates about knowledge production, postcolonial power, and the evolving digital humanities approaches that allow new ways to study both African media history and the processes of decolonization.

1. Decolonization and Information: Shaping the Terrain

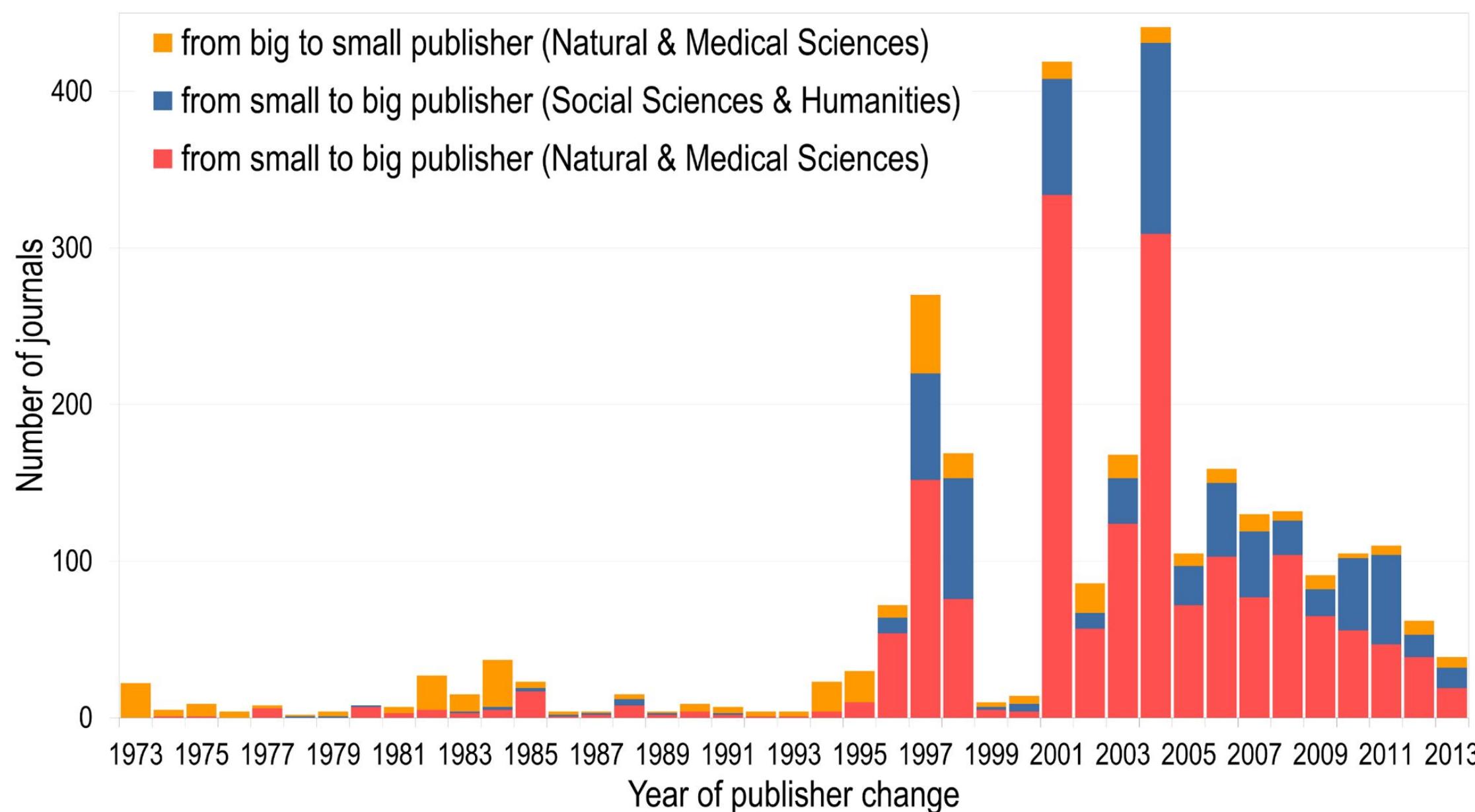
1.1. Early Histories of Colonial Knowledge Systems

Mid-twentieth-century historians of Africa, such as Kenneth Onwuka Dike (1956) and Jean Suret-Canale (1964), established how European colonial administrations not only seized economic and political control but also dominated the production and circulation of information. Dike's seminal *Trade and Politics in the Niger Delta, 1830–1885* (Clarendon Press) demonstrated how colonial records and archives shaped early historical narratives of African societies. These works underscored the fact that knowledge and "facts" about Africa were mediated by colonial offices, missionaries, and merchants, highlighting the structural imbalances that later African historians and media scholars would seek to redress.

Initial Results

Reference	Attempt 1	Attempt 2	Accurate?	Error	Previously Knew of Source?	Would Use Source?
Akyeampong, Emmanuel. 2000. "History, Memory, Slave-Trade and Slavery in Anlo (Ghana)." <i>Slavery & Abolition</i> 21(3): 1–24.	✓	✓	✗	Published in December 2001, not 2000	✗	😊
Bayart, Jean-François. 1993. <i>The State in Africa: The Politics of the Belly</i> . Longman.	✗	✓	✓		✗	😊
Brennan, James R. 2010. "Politics and Business in the Todd's Theatres of Colonial Tanganyika." <i>African Studies Review</i> 53(3): 1–24.	✓	✓	✗	Article doesn't exist		
Chakrabarty, Dipesh. 2000. <i>Provincializing Europe: Postcolonial Thought and Historical Difference</i> . Princeton University Press.	✓	✓	✓		✓	✓
Cooper, Frederick. 2002. <i>Africa Since 1940: The Past of the Present</i> . Cambridge University Press.	✓	✓	✓		✓	✓
Cooper, Frederick. 2014. <i>Africa in the World: Capitalism, Empire, Nation-State</i> . Harvard University Press.	✓	✓	✓		✓	✓
Dike, Kenneth Onwuka. 1956. <i>Trade and Politics in the Niger Delta, 1830–1885</i> . Clarendon Press.	✓	✓	✓		✗	✗
Falola, Toyin. 1993. <i>History of Nigeria 3: Nigeria in the Nineteenth Century</i> . Longman.	✓	✗	✓		✗	✗
Mbembe, Achille. 2001. <i>On the Postcolony</i> . University of California Press.	✓	✓	✓		✓	✓
Mbah, Emmanuel N. 2016. "Communication and Conflict Resolution in Postcolonial Cameroon." <i>African Studies Review</i> 59(2): 63–84.	✓	✗	✗	Article doesn't exist		
Mowlana, Hamid. 1980. "Communication in the Post-Industrial Era: New Trends in International Communication." <i>Journal of Communication</i> 30(2): 114–123.	✗	✓	✗	Article doesn't exist		
Ngũgĩ wa Thiong'o. 1986. <i>Decolonising the Mind: The Politics of Language in African Literature</i> . James Currey.	✓	✓	✓		✗	✓

Walled Gardens & Disciplinary Norms



Where are all these new academic tools coming from?



New mega citation indexes



Citation based literature mapping services



Science mapping tools



Research Graphs

CHR Forum

Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010

IAN MILLIGAN

*Abstract: It all seems so orderly and comprehensive. Instead of firing up the microfilm reader to navigate the *Globe and Mail* or the *Toronto Star*, one needs only to log into online newspaper databases. A keyword search, for a particular event, person, or cultural phenomenon, brings up a list of research findings. Previously impossible research projects can now be attempted. This process has fundamentally reshaped Canadian historical scholarship. We can phénomén see this in Canadian history dissertations. In 1998, a year with 67 dissertations, the *Toronto Star* was cited 74 times. However it was cited 753 times in 2010, a year with 69 dissertations. Similar data appears in the *Canadian Historical Review* (CHR), a prestigious peer-reviewed journal. Databases are skewing our research. We are witnessing the application of commercial Optical Character Recognition (OCR) technology – originally and primarily designed for the efficient digitization of large reams of corporate and legal documents, conventionally formatted – to historical sources. The results are, unsurprisingly, a mixed bag. In this article, I make two arguments. Firstly, online historical databases have profoundly shaped Canadian historiography. In a shift that is rarely – if ever – made explicit, Canadian historians have profoundly reacted to the availability of online databases. Secondly, historians need to understand how OCR works, in order to bring a level of methodological rigor to their work that use these sources.*

Keywords: historiography, digital history, databases, newspapers, primary sources, historical methodology, pedagogy

CHR advance online article
doi: 10.3138/chr.694

Ian Milligan highlights this by comparing newspaper citations in Canadian history dissertations before and after the introduction of major digital databases.

- Toronto Star citations surged from 74 (1998) to 753 (2010), a 991% increase.
- Globe and Mail citations rose from 58 to 708, a 926% increase.
- Montreal Gazette remained relatively stable, with a 16% increase (136 to 162).
- Toronto Telegram citations declined by 72% (31 to 19).

Literature Review & Summarization

Welcome to NotebookLM
EXPERIMENTAL

Choose File

Document guide Suggest three new features for the product based on this marketing plan

Back to results Monarchs in a Changing World / 3 Monarch Habitat Conservation across No...

BOOK CHAPTER

3 Monarch Habitat Conservation across North America: Past Progress and Future Needs

Priya C. Shahani, Guadalupe del Río Pesado, Phil Schappert, Eligio García Serrano
Monarchs in a Changing World: Biology and Conservation of an Iconic Butterfly, 2015, pp. 31-42 (12 pages)

https://www.jstor.org/stable/10.7591/j.ctt20fw696.7

Download Save Share Cite Additional information

NotebookLM early access program

- Up to one hundred projects, each with a max
- Max 50,000 words per source

1 / 12

3

Monarch Habitat Conservation across North America
Past Progress and Future Needs

Priya C. Shahani, Guadalupe del Río Pesado, Phil Schappert, and Eligio García Serrano

Preserving monarchs and their remarkable migration requires protecting habitats crucial for

Elicit How effective is open education in Asia?

Paper title: Democratization of Education through Massive Open Online Courses in Asia
Author: R. P. Barger
Journal: IAFOR Journal of Education
Year: 2020 Review 4 Citations PDF

Add information about all papers
Takeaway from abstract ✓
Intervention
Outcomes measured ✓
Number of participants
Population summary

BETA JSTOR AI research tool ON

Hi! I'm an AI-powered research tool designed to help you work more efficiently. I can summarize a text to help you quickly assess its relevance, recommend related JSTOR content, suggest thematically similar topics to further your research, and answer questions about what you're reading.

You: How is "history of monarch butterfly conservation efforts" related to this text?

The history of monarch butterfly conservation efforts includes several key initiatives and collaborations across North America:

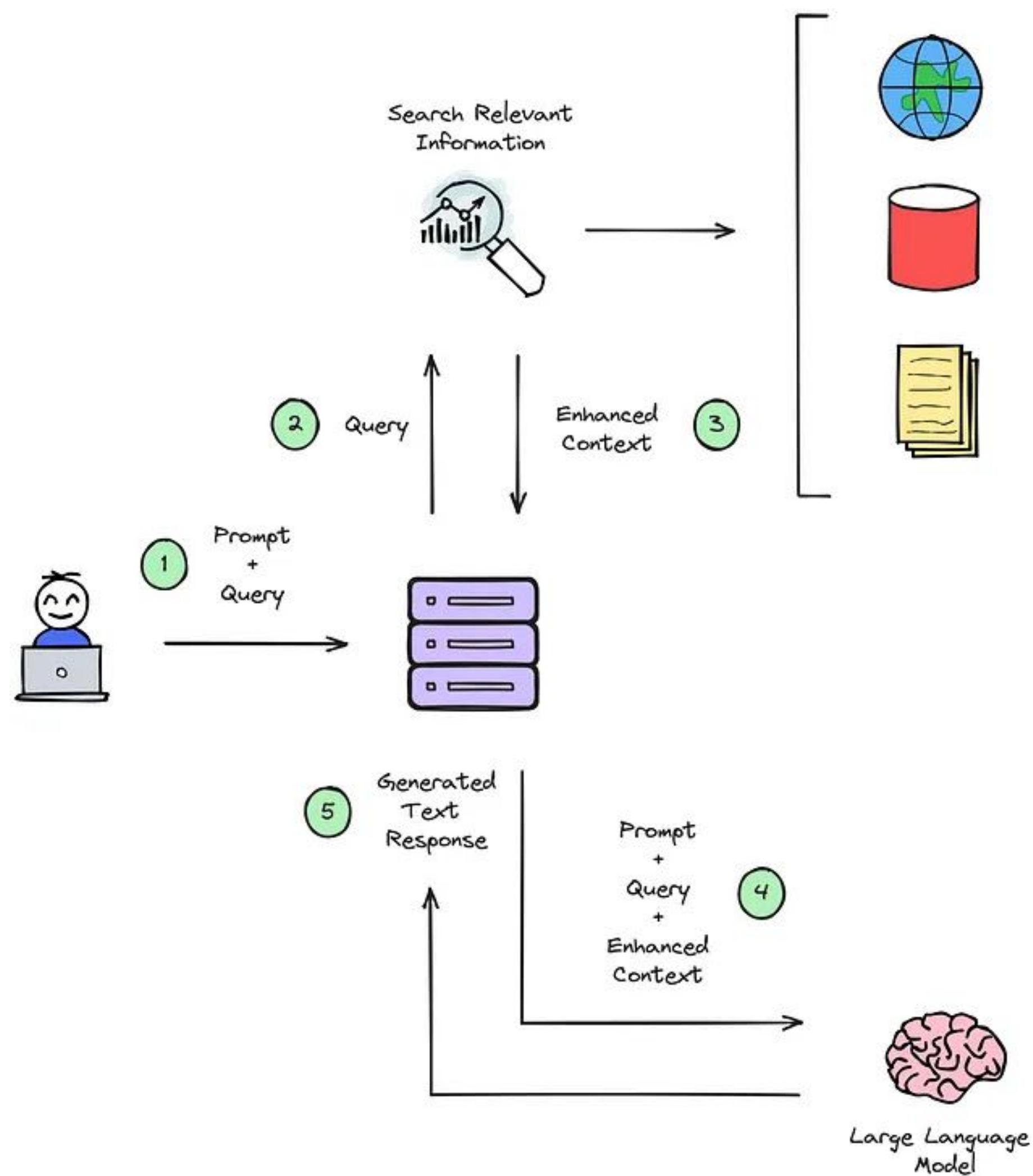
What is this text about? Recommend topics
Show me related content Ask a question about this text
Ask a question Send

We are continuously improving this tool with your feedback. While responses are based on content on JSTOR, the tool may produce biased or inaccurate results. [Learn more](#).

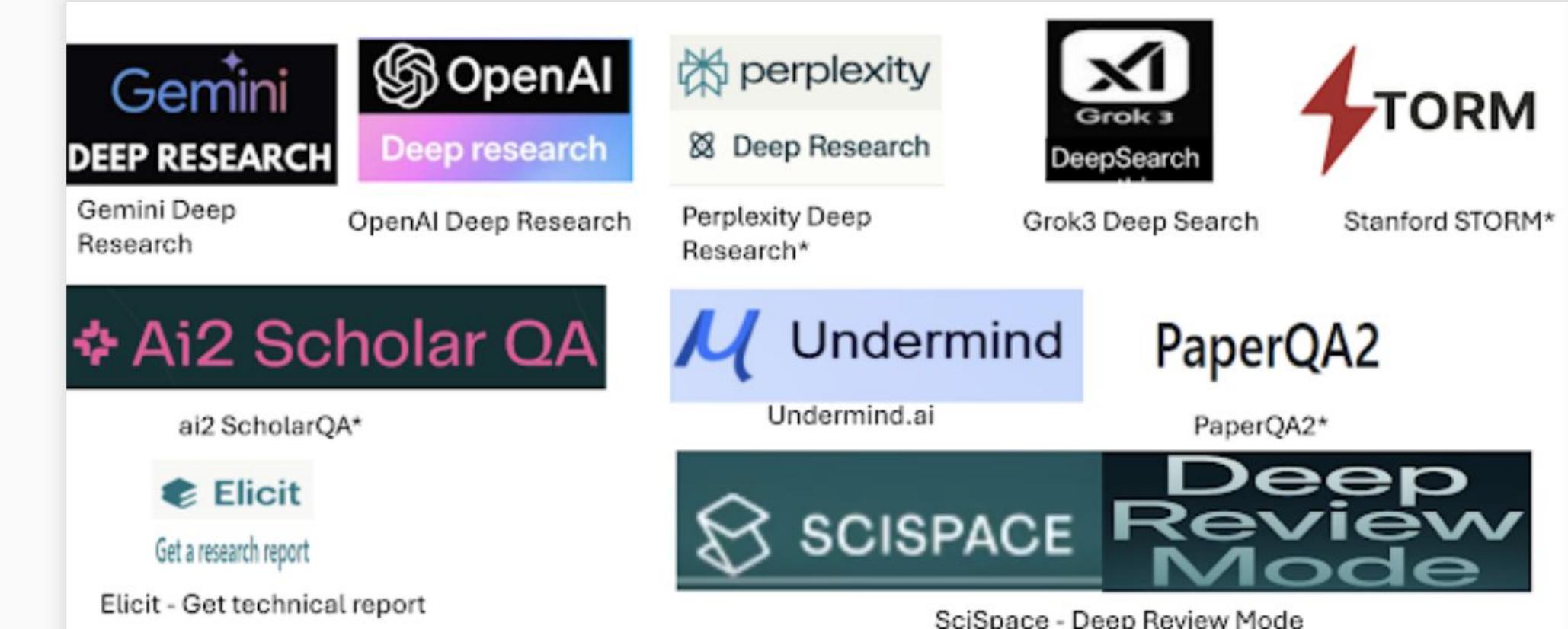
Submit feedback

This topic will be covered more in-depth with Jonathan Barber's breakout group on Literature Reviews & AI

Core Concepts With AI: Retrieval Augmented Generation



FEB
20
The Rise of Agent-Based Deep Research: Exploring OpenAI's Deep Research, Gemini Deep Research, Perplexity Deep Research, Ai2 ScholarQA, STORM, and More in 2025



When To Use AI? Trust & Validation

15 Times to use AI, and 5 Not to

Notes on the Practical Wisdom of AI Use



ETHAN MOLICK
DEC 09, 2024

930 48 102

Share

There are several types of work where AI can be particularly useful, given the current capabilities and limitations of LLMs. Though this list is based in science, it draws even more from experience. Like any form of wisdom, using AI well requires holding opposing ideas in mind: it can be transformative yet must be approached with skepticism, powerful yet prone to subtle failures, essential for some tasks yet actively harmful for others. I also want to caveat that you shouldn't take this list too seriously except as inspiration - you know your own situation best, and local knowledge matters more than any general principles. With all that out of the way, below are several types of tasks where AI can be especially useful, given current capabilities—and some scenarios where you should remain wary.

Aaron Tay's Musings about librarianship (Subscribe by email)

search

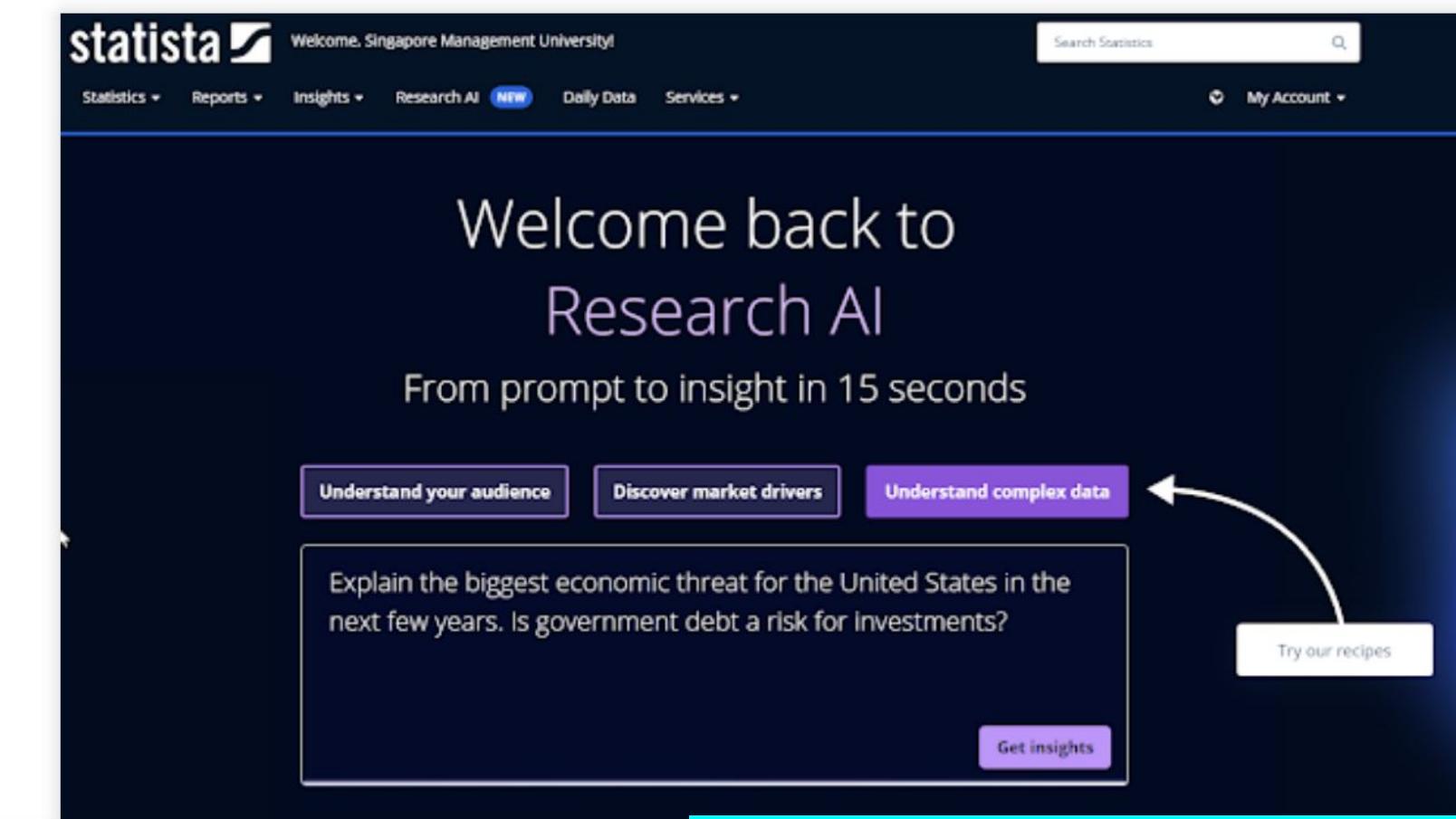
Magazine ▾ | About me Subscribe by email Best of posts series List of innovative search tools So you want to be a academic librarian in Singapore?

JUL
13

Prompt engineering with Retrieval Augmented Generation systems - tread with caution!

New! Listen to a 7 min autogenerated podcast discussing this blogpost from Google LLM Notebook

I recently watched a librarian give a talk about their experiments teaching prompt engineering. The librarian drawing from the academic literature on the subject (there are lots!), tried to leverage "prompt engineering principles" from one such paper to craft a prompt and used it in a Retrieval Augmented Generation (RAG) system, more specifically, Statista's brand new "research AI" feature.



If you need 100% accuracy, use AI tools & models carefully

Part 4

**Designing & Analyzing
Research Results
With AI**

Working with AI: Data Extraction

The leading AI models are now good historians

... in specific domains. Three case studies with GPT-4o, o1, and Claude Sonnet 3.5, and what they mean



BENJAMIN BREEN
JAN 22, 2025

107

30

Descritta da Urbano Monte Gentil'huomo Milanese, et dall'istesso disegnata in sessanta Tavole, Graduate in Longhezza, et Larghezza, conforme alla Geografia de molti veridici, si antichi, come moderni autori. Con sue esposizioni a tutte le Tavole, particolarmente della qualità de i paesi, et delle cose più notabili, che in essi si ritrovano.
Quale Tavole, si posson, et s'insegnano a ridurre insieme, et di quelle = unirne, et formarne un Mappamondo intero in forma piana circolare!
Auanti alle quale, per maggior intelligenza, et satisfactione de Studenti, vi sono posti doi libri, de quali il primo Tratta de molte cose naturali concernenti alla qualità del Mondo, Et l'altro si distingue con pieno discorso, de tutti i Termini, et regole di Geografia.
Con una copiosa Tavola de nomi antichi, de Provincie, città, Mari, Fiumi, et altre cose dechiarati con i moderni.
E' molte altre Tavole, et ruote di bellissima speculazione, et necessarie, che leggendo ciascuno potra conoscere?
Finalmente doppo alcune cose intorno a gli ecclissi del sole, et della Luna, et alcuni loro effetti, et alcune altre intorno alla qualità, et quantità de i venti, et come si facia il Bussolo da navigare, et modo d'intendere, et adoperare la carta da navigare: vi si pongono doi Indici particolari di questo libro, l'uno sull' Tavole et Capitoli, et l'altro delle cose contenute nel libro più notabili.

Descritto da Urbano Monte Gentil'huomo Milanese, et dall'istesso disegnato in sessanta Tavole, Graduate in Longhezza, et Larghezza, conforme alla Geografia de molti veridici, si antichi, come moderni autori. Con sue esposizioni a tutte le Tavole, particolarmente della qualità de' paesi, et delle cose più notabili, che in essi si ritrovano.
Quale Tavole, si possono, et s'insegnano a ridurre insieme, et de quelle venire, et formarne un Mappamondo intero in forma piana Circolare. Avanti alle quale, per maggior intelligenza, et satisfactione de Studenti, vi sono posti doi Libri, de quali il primo tratta de molte cose naturali concernenti alla qualità del Mondo, Et l'altro si distingue con pieno discorso, de tutti i Termini, et regole di Geografia.
Con una copiosa Tavola de nomi antichi, de Provincie, città, Mari, Fiumi, et altre cose dechiarati con i moderni.
Et molte altre Tavole, et note di bellissima speculazione, et necessarie, che leggendo ciascuno potrà conoscere.
Finalmente doppo alcune cose intorno a gli ecclissi del sole, et della Luna, et alcuni loro effetti, et alcune altre intorno alla qualità, et quantità de i venti, et come si tracia il Bussolo da navigare, et modo d'intendere, et adoperare la carta da navigare: vi si pongono doi Indici particolari di questo libro, uno delle Tavole et Capitoli, et l'altro delle cose contenute nel libro più notabili.

Working with AI: Data Extraction

Home / MSU Libraries Digital Collections / Africana Studies Collection / African e-Journals Project / Africa Media Review

A critical appraisal of the communication gap in the liberation struggle

This paper calls for a critical appraisal of the use of the politically loaded words which are normally associated with liberation struggles. It suggests that these terms are often vaguely defined and mean different things to different publics. Their constant use, therefore, tends to obscure the message and to create a communication gap. The writer subscribes to the view that, contrary to the intention of the message senders, many of the descriptors used present a denigrating image of the groups they set out to help and that both these descriptors as
[Read more](#)

In Collections Africa Media Review

Copyright Status In Copyright

Date Published 1987

Authors Sybil, L. James

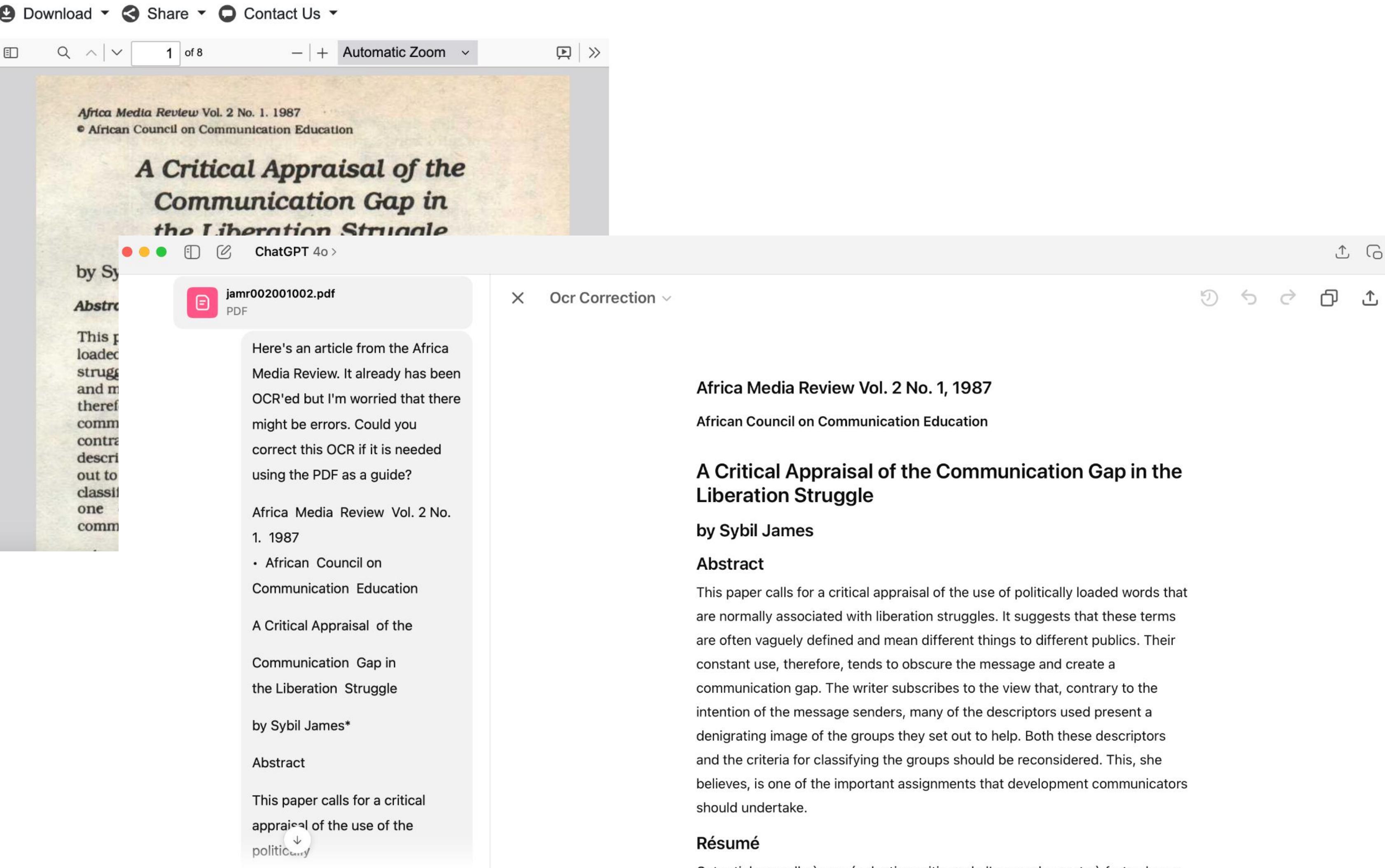
Subjects Communication in politics ([More info](#))

Reference (Linguistics) ([More info](#))

Language and languages--Political aspects ([More info](#))

Material Type Articles

Publishers Institute for Communication Development and Research (African Council on Communication Education)



The screenshot shows a Mac OS X desktop environment. A PDF document titled "A Critical Appraisal of the Communication Gap in the Liberation Struggle" by Sybil James is open in a browser window. The document is from the "Africa Media Review Vol. 2 No. 1. 1987" and is published by the "African Council on Communication Education". A floating window titled "ChatGPT 4.0" contains the text: "Here's an article from the Africa Media Review. It already has been OCR'd but I'm worried that there might be errors. Could you correct this OCR if it is needed using the PDF as a guide?". Below this, the document's abstract is visible, stating: "This paper calls for a critical appraisal of the use of politically loaded words that are normally associated with liberation struggles. It suggests that these terms are often vaguely defined and mean different things to different publics. Their constant use, therefore, tends to obscure the message and create a communication gap. The writer subscribes to the view that, contrary to the intention of the message senders, many of the descriptors used present a denigrating image of the groups they set out to help. Both these descriptors and the criteria for classifying the groups should be reconsidered. This, she believes, is one of the important assignments that development communicators should undertake." The ChatGPT window also includes a file icon and the path "jamr002001002.pdf". The overall interface includes standard Mac OS X window controls and a toolbar at the top.

Limitations of Chatbot UI & Context Windows

Context window

Models on this page will list a **context window**, which refers to the maximum number of tokens that can be used in a single request, inclusive of both input, output, and reasoning tokens. For example, when making an API request to [chat completions](#) with the [o1 model](#), the following token counts will apply toward the context window total:

- Input tokens (inputs you include in the `messages` array with [chat completions](#))
- Output tokens (tokens generated in response to your prompt)
- Reasoning tokens (used by the model to plan a response)

Tokens generated in excess of the context window limit may be truncated in API responses.

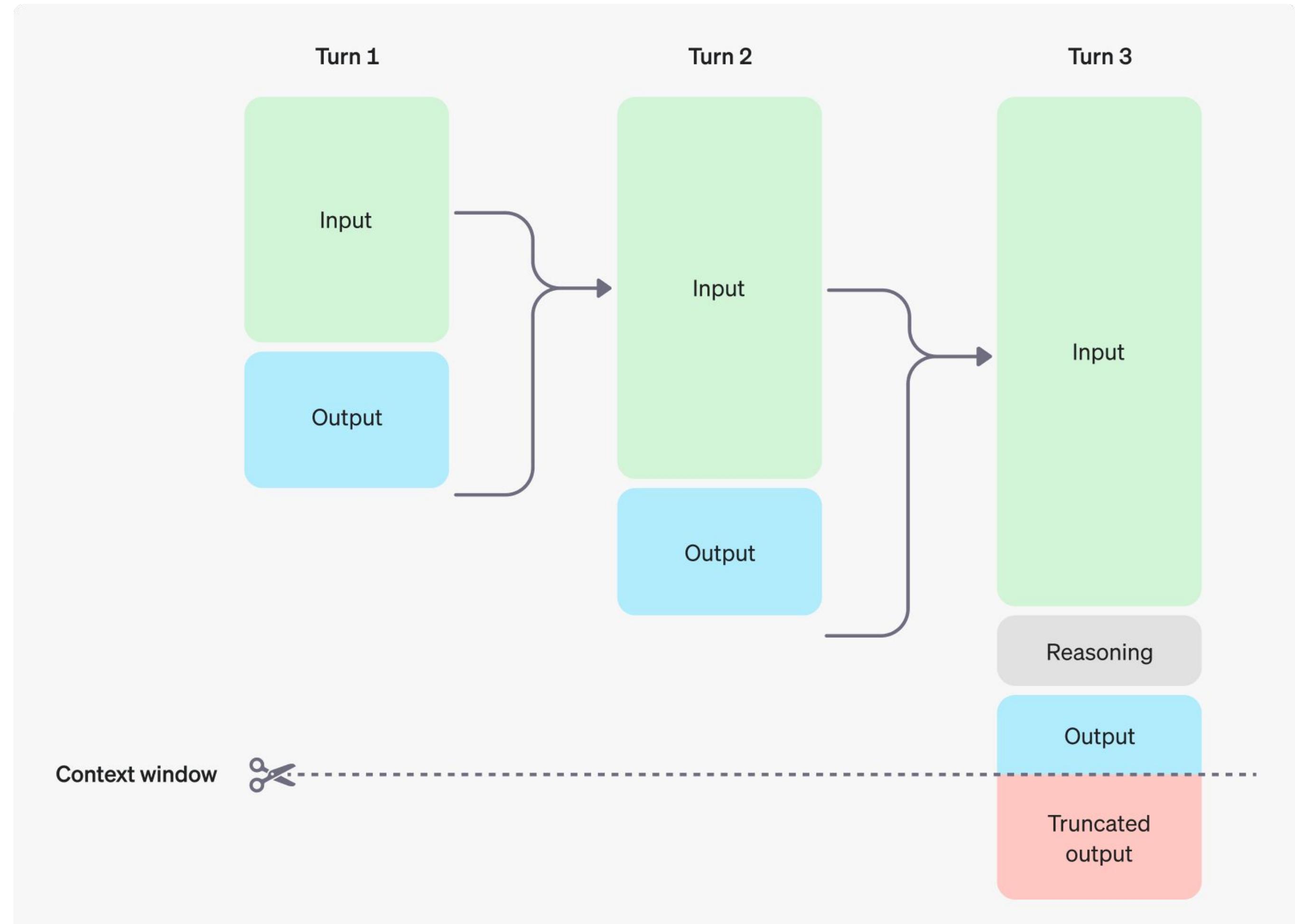
GPT-4o

GPT-4o (“o” for “omni”) is our versatile, high-intelligence flagship model. It accepts both text and [image inputs](#), and produces [text outputs](#) (including [Structured Outputs](#)). Learn how to use GPT-4o in our [text generation guide](#).

The `chatgpt-4o-latest` model ID below continuously points to the version of GPT-4o used in [ChatGPT](#). It is updated frequently, when there are significant changes to ChatGPT's GPT-4o model.

The knowledge cutoff for GPT-4o models is **October, 2023**.

MODEL	CONTEXT WINDOW	MAX OUTPUT TOKENS
gpt-4o ↳ gpt-4o-2024-08-06	128,000 tokens	16,384 tokens
gpt-4o-2024-11-20	128,000 tokens	16,384 tokens
gpt-4o-2024-08-06	128,000 tokens	16,384 tokens
gpt-4o-2024-05-13	128,000 tokens	4,096 tokens
chatgpt-4o-latest ↳ GPT-4o used in ChatGPT	128,000 tokens	16,384 tokens



Working with AI: Getting Data Programmatically

ChatGPT 4o >

This topic will be covered more in-depth in my breakout group on Data Analysis & AI

I want to download all the data for the Africa Media Review from the MSU digitized collections. There's 293 separate web pages and each has an article with metadata, pdf, and extracted OCR that ties back to an issue. I could do this manually but it's very tedious. How would suggest I try to access and collect these materials?

To automate the collection of all **Africa Media Review** data from the **MSU digitized collections**, I'd suggest a web scraping approach. Here's a step-by-step strategy:

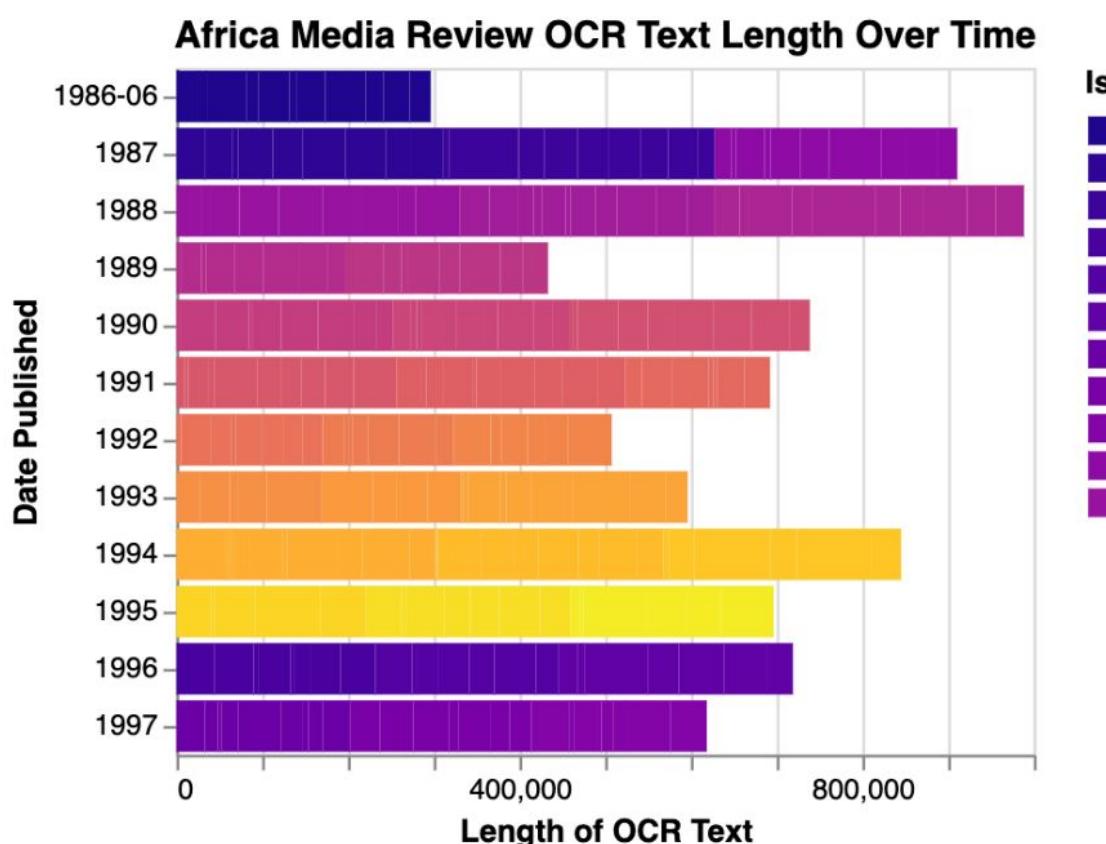
1. Investigate the Site's Structure

- Browse a few of the **293 web pages** and check:
 - URL patterns (e.g., <https://example.edu/africa-media-review/issue-xx/article-yy>)
 - Structure of the metadata, PDF links, and OCR text.
 - Whether pagination or issue-based indexing is used.

• Use your browser's **Developer Tools (Inspect Element)** to examine:

- HTML structure (tags for metadata, OCR, PDF links)
- API calls (if the site dynamically loads data via JavaScript)

The screenshot shows a web browser window with the URL d.lib.msu.edu/search?sort=title%20asc&fq=RELS_EXT_isMemberOfCollection_uri_s%3Ainfo%5C3Afedora/jamr%5C.... The page displays a search interface with filters for 'Collection: Africa Media Review' and 'Material Type'. Below the filters, there is a 'Search results' section showing 20 of 293 results. One result is highlighted, showing a thumbnail of a document titled 'A critical appraisal of the communication gap in the liberation struggle' by Sybil, L. James, published in Text (1987). The bottom of the screenshot shows the browser's developer tools, specifically the Elements tab, displaying the HTML code for a search result item.



Issue

- Africa media review. Vol. 1 no. 1 ...
- Africa media review. Vol. 1 no. 2 ...
- Africa media review. Vol. 1 no. 3 ...
- Africa media review. Vol. 10 no. 1 ...
- Africa media review. Vol. 10 no. 2 ...
- Africa media review. Vol. 10 no. 3 ...
- Africa media review. Vol. 11 no. 1 ...
- Africa media review. Vol. 11 no. 2 ...
- Africa media review. Vol. 11 no. 3 ...
- Africa media review. Vol. 2 no. 1 ...
- Africa media review. Vol. 2 no. 2 ...
- Africa media review. Vol. 2 no. 3 ...
- Africa media review. Vol. 3 no. 1 ...
- Africa media review. Vol. 3 no. 2 ...
- Africa media review. Vol. 3 no. 3 ...
- Africa media review. Vol. 4 no. 1 ...
- Africa media review. Vol. 4 no. 2 ...
- Africa media review. Vol. 4 no. 3 ...
- Africa media review. Vol. 5 no. 1 ...
- Africa media review. Vol. 5 no. 2 ...
- Africa media review. Vol. 5 no. 3 ...
- Africa media review. Vol. 6 no. 1 ...
- Africa media review. Vol. 6 no. 2 ...
- Africa media review. Vol. 6 no. 3 ...
- Africa media review. Vol. 7 no. 1 ...
- Africa media review. Vol. 7 no. 2 ...
- Africa media review. Vol. 7 no. 3 ...
- Africa media review. Vol. 8 no. 1 ...
- Africa media review. Vol. 8 no. 2 ...
- Africa media review. Vol. 8 no. 3 ...
- Africa media review. Vol. 9 no. 1 ...
- Africa media review. Vol. 9 no. 2 ...
- Africa media review. Vol. 9 no. 3 ...

Limitations of Coding with Chatbots

A screenshot of a Jupyter Notebook interface. The title bar says "Web Scraping Workshop". The code cell contains Python code for web scraping, specifically extracting metadata from a single article. A modal dialog box is overlaid on the screen, asking "Allow network access?". The dialog provides a brief explanation: "'Web Scraping Workshop' is trying to connect to the network. If you don't recognize this request, don't allow this app access to the network." It includes a URL: https://d.lib.msu.edu/search?sort=title%20asc&fq=RELS_EX. At the bottom of the dialog are two buttons: "Deny" and "Allow", with "Allow" being highlighted. Below the dialog, the notebook shows the execution of the code, with several warning messages in yellow text about Arrow string types, interoperability, and missing dependencies like `pyarrow`.

```
45 all_links_df = pd.DataFrame(all_links, columns=["collection_item_link"])
46
47 # Step 5: Extract Metadata from One Article
48 # Let's scrape a single article to understand the data structure.
49
50 sample_url = all_links_df.iloc[0]["collection_item_link"]
51 response = requests.get(sample_url)
52 soup = BeautifulSoup(response.content, "html.parser")
53
54 # Extract title
55 title_element = soup.find("h1")
56 title = title_element.text if title_element else "No title found"
57
58 # Extract abstract
59 abstract_element = soup.find("div", class="abstract")
60 abstract = abstract_element.text if abstract_element else "No abstract found"
61
62 # Extract OCP
63 transcript_url = transcript_re.match(title).group("url")
64 transcript_response = requests.get(transcript_url)
65 transcript_text = transcript_response.text
66 transcript_te
67 # Extract metadata
```

Console

Run running code ...

[warn]: (to allow more performant data types, such as the Arrow string type, and better interoperability with other libraries)

[warn]: but was not found to be installed on your system.

[warn]: If this would cause problems for you,

[warn]: please provide us feedback at <https://github.com/pandas-dev/pandas/issues/54466>

[warn]:

INTERDISCIPLINARY SCIENCE REVIEWS, Vol. 40 No. 4, December 2015, 376–392

Trusting Others to ‘Do the Math’

REBECCA SUTTON KOESER

Emory University Libraries & IT Services, Emory University, Atlanta, Georgia, USA

Researchers effectively trust the work of others anytime they use software tools or custom software. In this article I explore this notion of trusting others, using Digital Humanities as a focus, and drawing on my own experience. Software is inherently flawed and limited, so when its use in scholarship demands better practices and terminology, to review research software and describe development processes. It is also important to make research software engineers and their work more visible, both for the purposes of review and credit.

KEYWORDS software development, research software engineers, digital publishing, digital humanities

Working with LLMs

The Prompt Report: A Systematic Survey of Prompting Techniques

Sander Schulhoff¹, Michael Ilie¹, Nishant Balepur¹, Konstantine Kahadze¹,

Amanda Liu¹, Chenglei Si³, Yinheng Li⁴, Aayush Gupta¹, Hyojung Han¹,

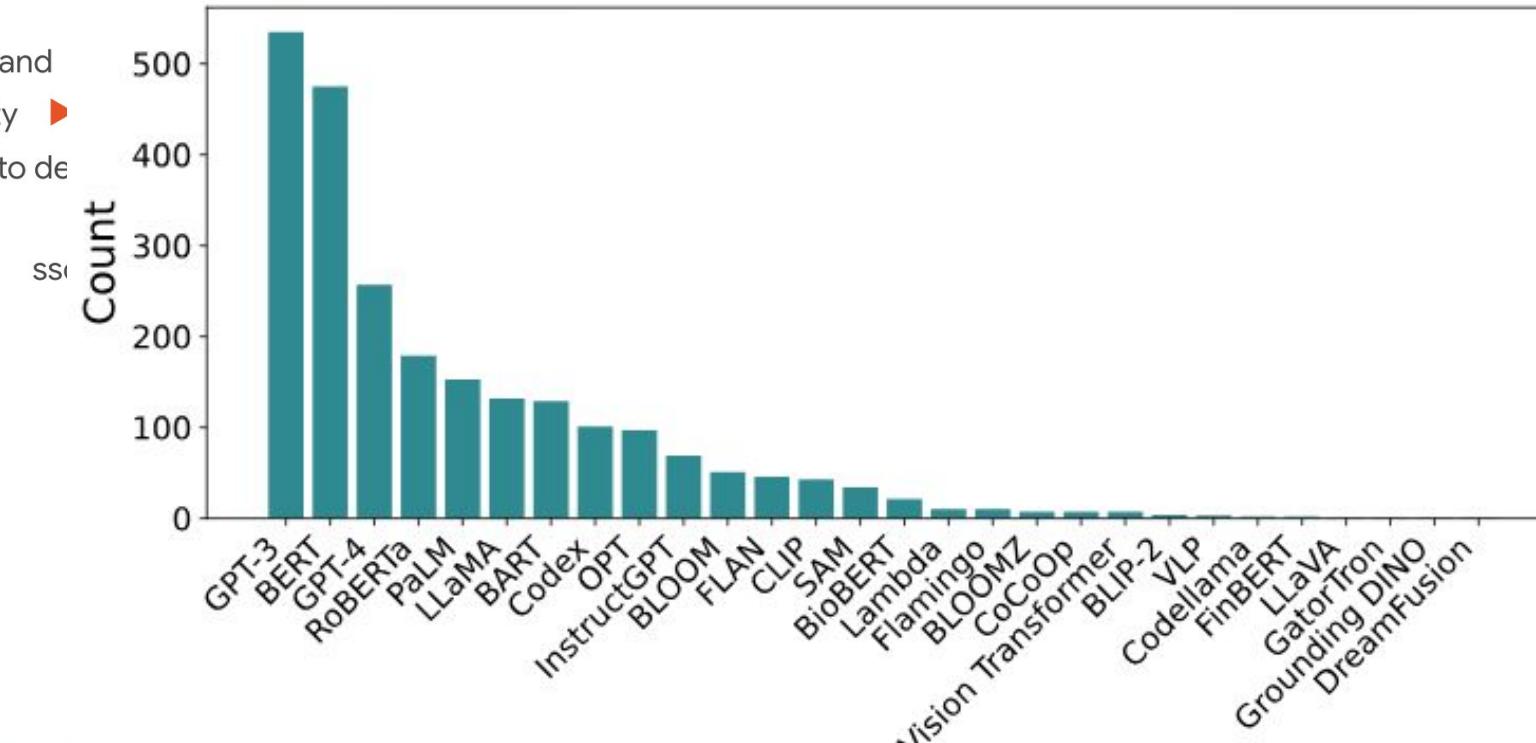
Sevien Schulhoff¹, Pranav Sandeep Dulepet¹, Saurav Vidyadhara¹, Dayeon Ki¹, Sweta Agrawal¹¹,

Chau Pham¹², Gerson Kroiz, Feileen Li¹, Hudson Tao¹, Ashay Srivastava¹, Hevander Da Costa¹,

Saloni Gupta¹, Megan L. Rogers⁷, Inna Gonccarenco⁸, Giuseppe Sarli^{8,9}, Igor Galynker¹⁰,

Denis Peskoff⁶, Marine Ca

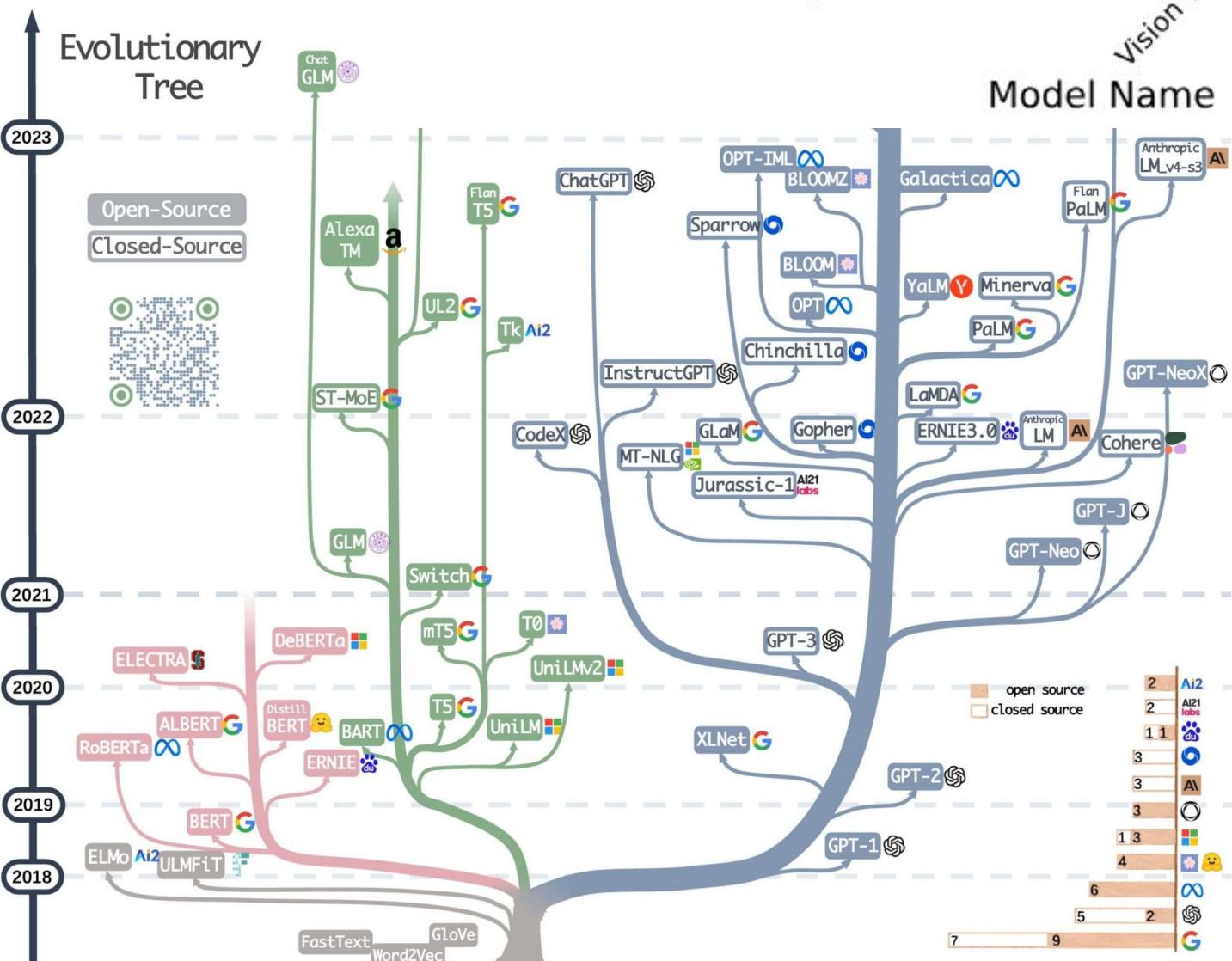
Counts of Model Mentions in Dataset



►¹ University of Maryland

►⁷ Texas State University

►¹¹ Instituto de



1. AI-Powered Text Summarization

- ◆ **Use Case:** Automatically summarize each transcript to get a sense of the content.
- ◆ **Tool:** OpenAI API, Hugging Face Transformers (T5, Bart, GPT-3.5)

Code Example

python

```
from transformers import pipeline

# Load summarization model
summarizer = pipeline("summarization")

# Summarize a sample transcript
sample_text = " ".join(transcripts_df.iloc[0]["transcript"].split()[:500]) # Limit to first 500 words
summary = summarizer(sample_text, max_length=100, min_length=30, do_sample=False)

print("Summary:", summary[0]['summary_text'])
```

Copy

Spaces: HuggingFaceH4/open_llm_leaderboard □ like 1.39k • Running

App Files Community 37

Open LLM Leaderboard

With the plethora of large language models (LLMs) and chatbots being released week upon week, often with grandiose claims of their performance, it can be hard to filter out the genuine progress that is being made by the open-source community and which model is the current state of the art. The 😊 Open LLM Leaderboard aims to track, rank and evaluate LLMs and chatbots as they are released.

😊 A key advantage of this leaderboard is that anyone from the community can submit a model for automated evaluation on the 😊 GPU cluster, as long as it is a 😊 Transformers model with weights on the Hub. We also support evaluation of models with delta-weights for non-commercial licensed models, such as LLaMa.

↗ In the first tab (LLM Benchmarks), we evaluate models on 4 key benchmarks from the [Eleuther AI Language Model Evaluation Harness](#), a unified framework to test generative language models on a large number of different evaluation tasks. In the second tab (Human & GPT Evaluations), TK.

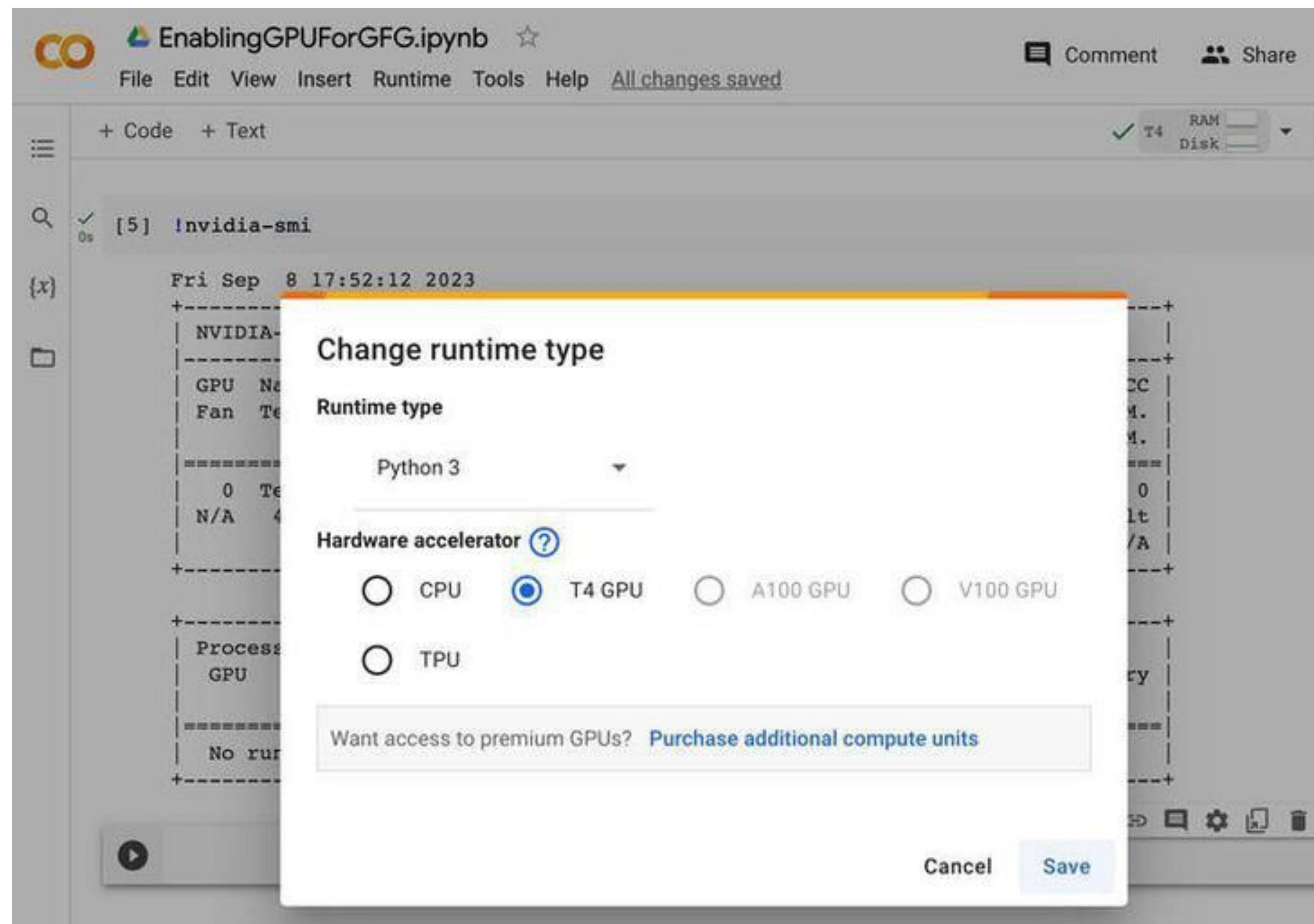
Evaluation is performed against 4 popular benchmarks:

- [AI2 Reasoning Challenge](#) (25-shot) - a set of grade-school science questions.
- [HellaSwag](#) (10-shot) - a test of commonsense inference, which is easy for humans (~95%) but challenging for SOTA models.
- [MMLU](#) (5-shot) - a test to measure a text model's multitask accuracy. The test covers 57 tasks including elementary mathematics, US history, computer science, law, and more.
- [TruthfulQA](#) (0-shot) - a benchmark to measure whether a language model is truthful in generating answers to questions.

We chose these benchmarks as they test a variety of reasoning and general knowledge across a wide variety of fields in 0-shot and few-shot settings.

Model	Revision	Average	ARC (25-shot)	HellaSwag (10-shot)	MMLU (5-s)
			↑	↑	↑
tiiuae/falcon-40b-instruct	main	63.2	61.6	84.4	54.1

Working with LLMs



```
EnablingGPUForGFG.ipynb
```

```
+ Code + Text
```

```
[5] !nvidia-smi
```

```
Fri Sep 8 17:52:12 2023
```

```
NVIDIA-
```

```
GPU N
```

```
Fan Te
```

```
0 Te
```

```
N/A 4
```

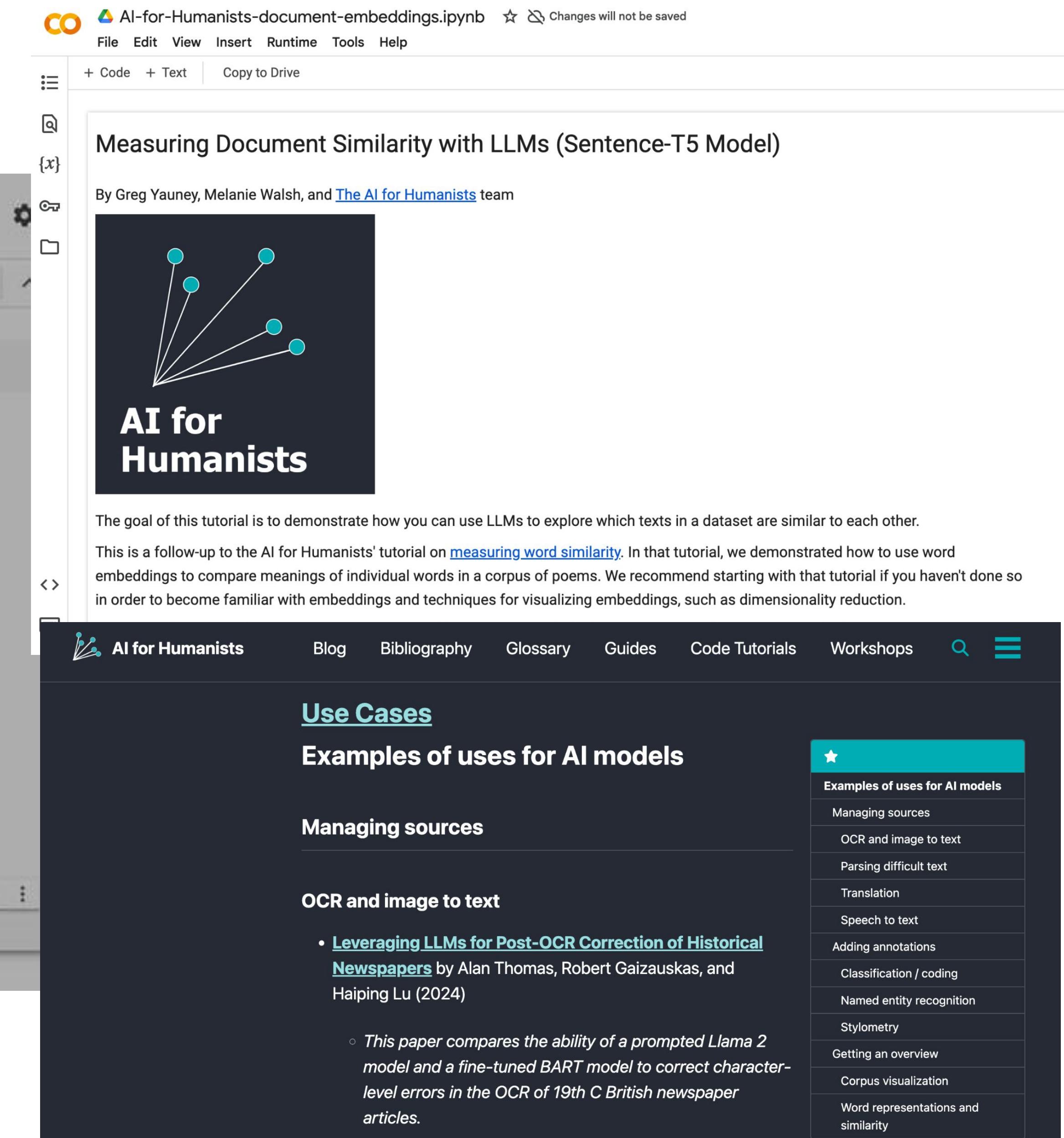
```
Process
```

```
GPU
```

```
No r
```

```
Want access to premium GPUs? Purchase additional compute units
```

```
Cancel Save
```



```
AI-for-Humanists-document-embeddings.ipynb
```

```
File Edit View Insert Runtime Tools Help
```

```
+ Code + Text Copy to Drive
```

Measuring Document Similarity with LLMs (Sentence-T5 Model)

By Greg Yauney, Melanie Walsh, and [The AI for Humanists](#) team



AI for Humanists

The goal of this tutorial is to demonstrate how you can use LLMs to explore which texts in a dataset are similar to each other. This is a follow-up to the AI for Humanists' tutorial on [measuring word similarity](#). In that tutorial, we demonstrated how to use word embeddings to compare meanings of individual words in a corpus of poems. We recommend starting with that tutorial if you haven't done so in order to become familiar with embeddings and techniques for visualizing embeddings, such as dimensionality reduction.

[AI for Humanists](#) Blog Bibliography Glossary Guides Code Tutorials Workshops

Use Cases

Examples of uses for AI models

Managing sources

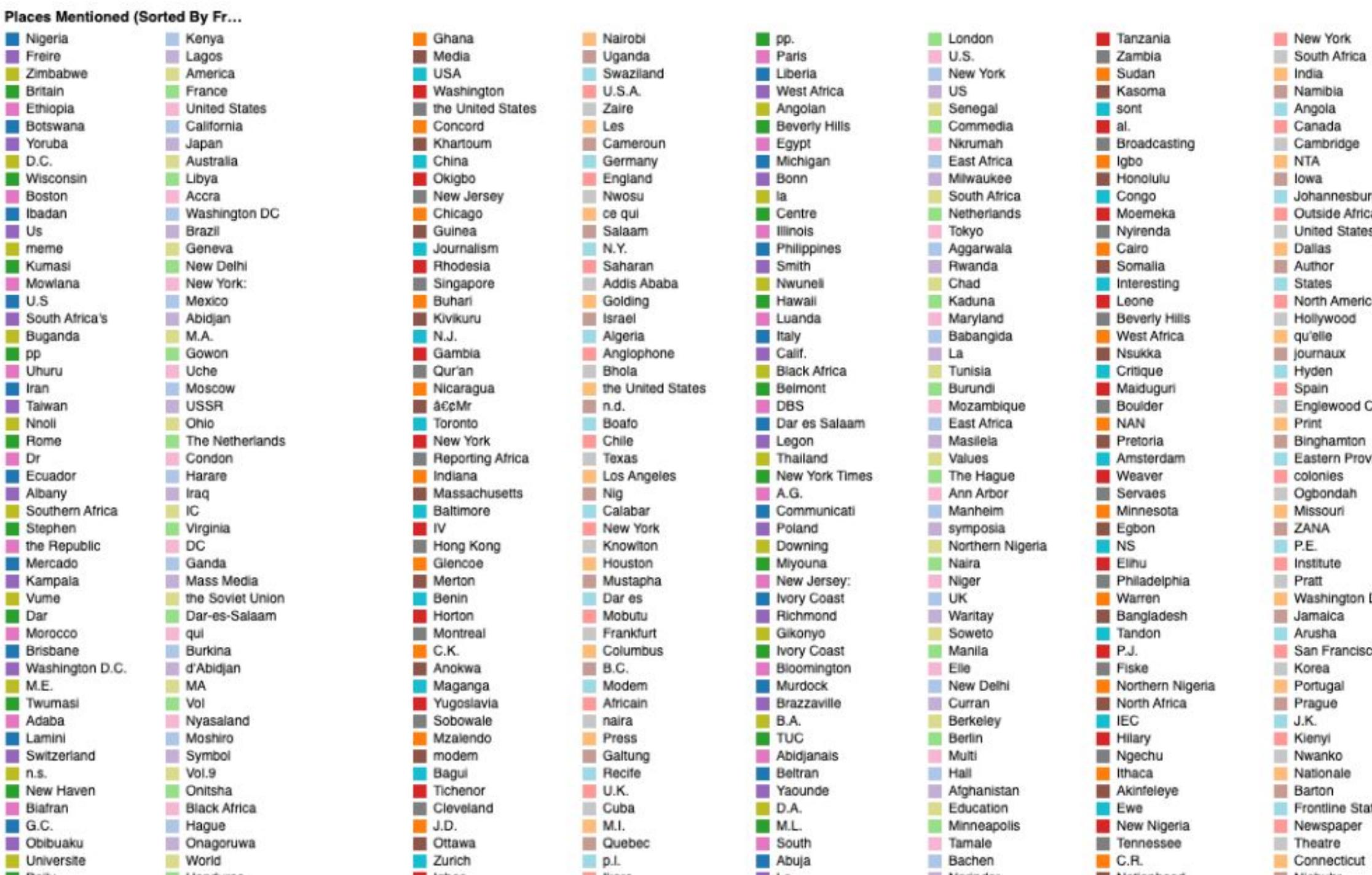
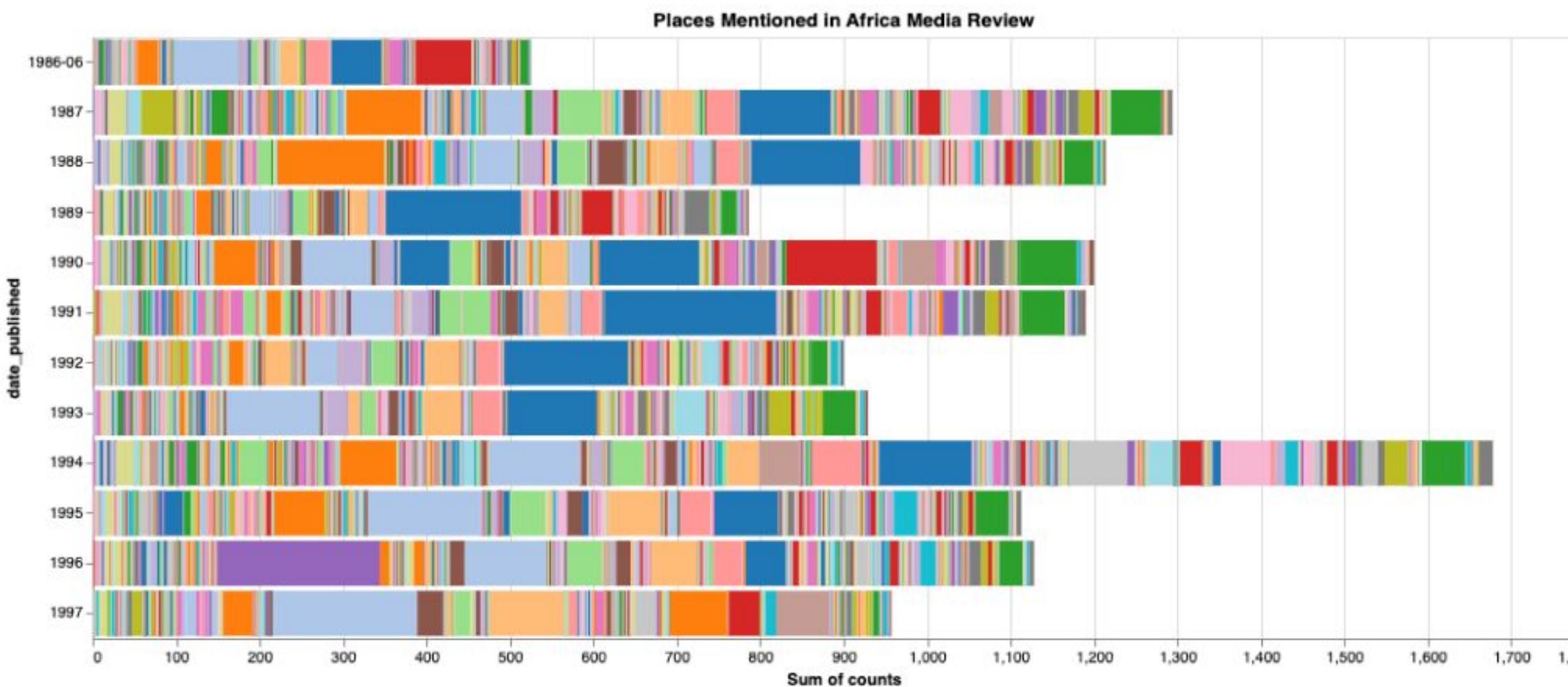
OCR and image to text

- [Leveraging LLMs for Post-OCR Correction of Historical Newspapers](#) by Alan Thomas, Robert Gaizauskas, and Haiping Lu (2024)
 - This paper compares the ability of a prompted Llama 2 model and a fine-tuned BART model to correct character-level errors in the OCR of 19th C British newspaper articles.*

★ Examples of uses for AI models

- Managing sources
- OCR and image to text
- Parsing difficult text
- Translation
- Speech to text
- Adding annotations
- Classification / coding
- Named entity recognition
- Styliometry
- Getting an overview
- Corpus visualization
- Word representations and similarity

Initial Results With LLMs



Reproducibility & Disciplinary Norms

SPRINGER LINK

Log in

Find a journal Publish with us Track your research

Search

Cart

Home > Collection

Reproducibility and Explainability in Digital Humanities

Participating journal: International Journal of Digital Humanities

Closed for submissions

"Reproducibility" and "explainability" are important methodological considerations in the Sciences, and are increasingly relevant in Digital Humanities. The discussions around Nan Z. Da's *The Computational Case against Computational Literary Studies* (2019), Katherine Bode's *Why You Can't Model Away Bias* (2020), Beatrice Fazi's *Beyond Human: Deep Learning, Explainability and Representation* (2020) and *Perceptual Bias and Technical Metapictures* (2020) by Fabian Offert and Peter Bell foregrounded a crucial methodological desideratum for the interconnection of digital methods, humanities research, and critical approaches. Many methods and tools prevalent in Digital Humanities research are not reproducible. With the advent of AI, the predictions of many models are not explainable. Finally, Digital Humanities research still has to develop standards, as well as a...

Show more

Participating journal



Journal

International Journal of Digital Humanities

The International Journal of Digital Humanities is a peer-reviewed academic journal with a focus on digital media and the development, application and reflection of digital research...

Publishing model	Hybrid
Journal Impact Factor	Not available
Downloads	55.8k
Submission to first decision	65 days (median)

Editors



Thorsten Ries

University of Texas, Austin, USA Dr Ries joined the Department of Germanic Studies in 2021 as Assistant Professor with specialism in German literature from the 18th to the 21st century, digital learning and digital...

Show more



Karina van Dalen-Oskam

University of Amsterdam, the Netherlands Karina van Dalen-Oskam studied Dutch Linguistics and Literary studies at the University of Utrecht and in 1997 gained her PhD from Leiden University. She is head of the...



The Turing Way

Foreword >

Guide for Reproducible Research

Overview of Reproducible Research >

Open Research >

Version Control >

Licensing >

Research Data Management >

Reproducible Environments >

BinderHub >

Code documentation >

Code Quality >

Code Testing >

Code Reviewing Process >

Reusable Code >

Continuous integration >

Reproducibility with Make >

Guide for Reproducible Research

The Turing Way Community

This guide covers topics related to skills, tools and best practices for research reproducibility.

The *Turing Way* defines reproducibility in data research as data and code being available to fully rerun the analysis.

There are several definitions of reproducibility in use, and we discuss these in more detail in the Definitions section of this chapter. While it is absolutely fine for us each to use different words, it will be useful for you to know how *The Turing Way* defines *reproducibility* to avoid misunderstandings when reading the rest of the handbook.

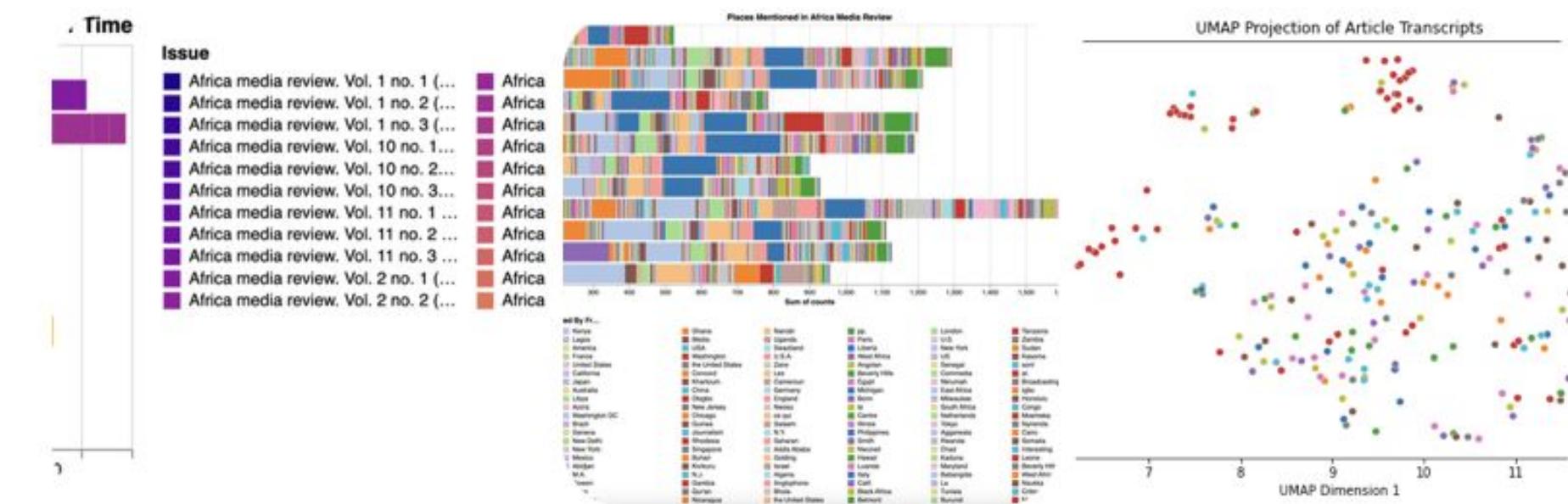


Figure 1: *The Turing Way* project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: [The Turing Way Community & Scriberia \(2024\)](#).

Part 5

**Evaluating & Preserving
Research With AI**

Final Results



I've now done some initial data analysis of the web scraped OCR data from the *Africa Media Review* issues available from the MSU digital collections. I've created a basic overview graph of the length of articles over time, run spaCy's NER model to extract place names, and then used the sentence transformer model and projected all the articles with UMAP. Can you interpret these graphs and then work them into your essay?

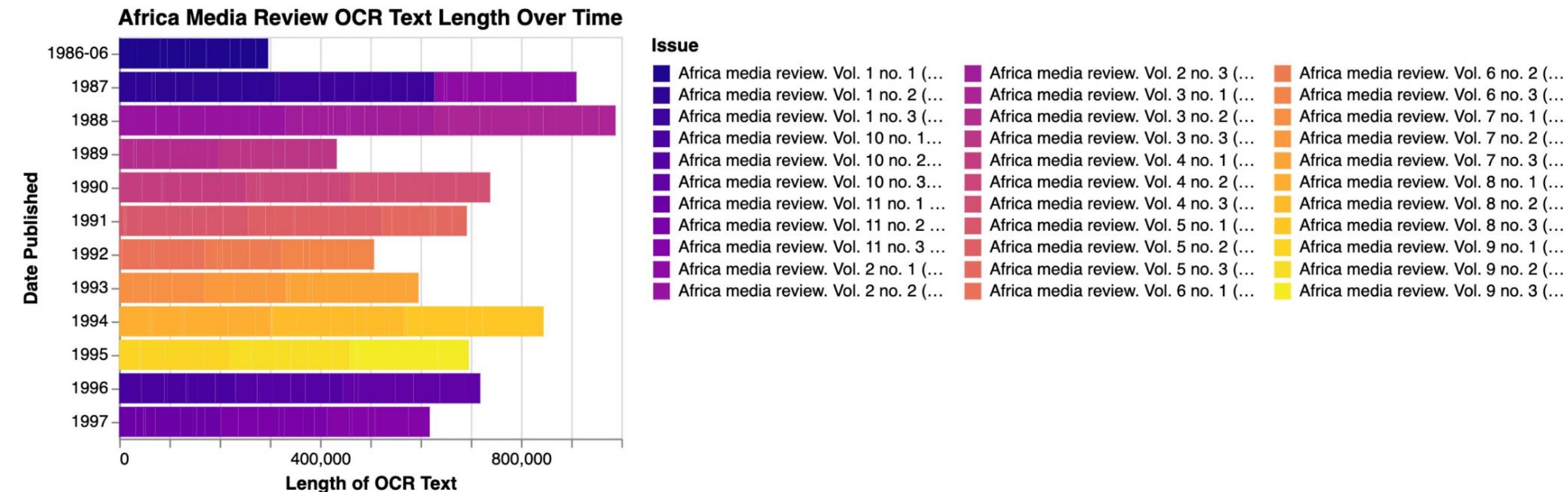


Request for o1 pro mode

Finished reasoning

Below is an expanded historiographical essay on *Africa Media Review* that integrates recent digital-humanities analyses of the journal's corpus—specifically, a bar chart illustrating overall OCR text length by publication date, a place-name frequency chart derived from named-entity recognition (NER), and a two-dimensional visualization (via UMAP) of the semantic relationships among articles. These graphical insights enrich our understanding of how *Africa Media Review* evolved over time, what geographic focal points animated its pages, and how its articles cluster thematically.

Final Results



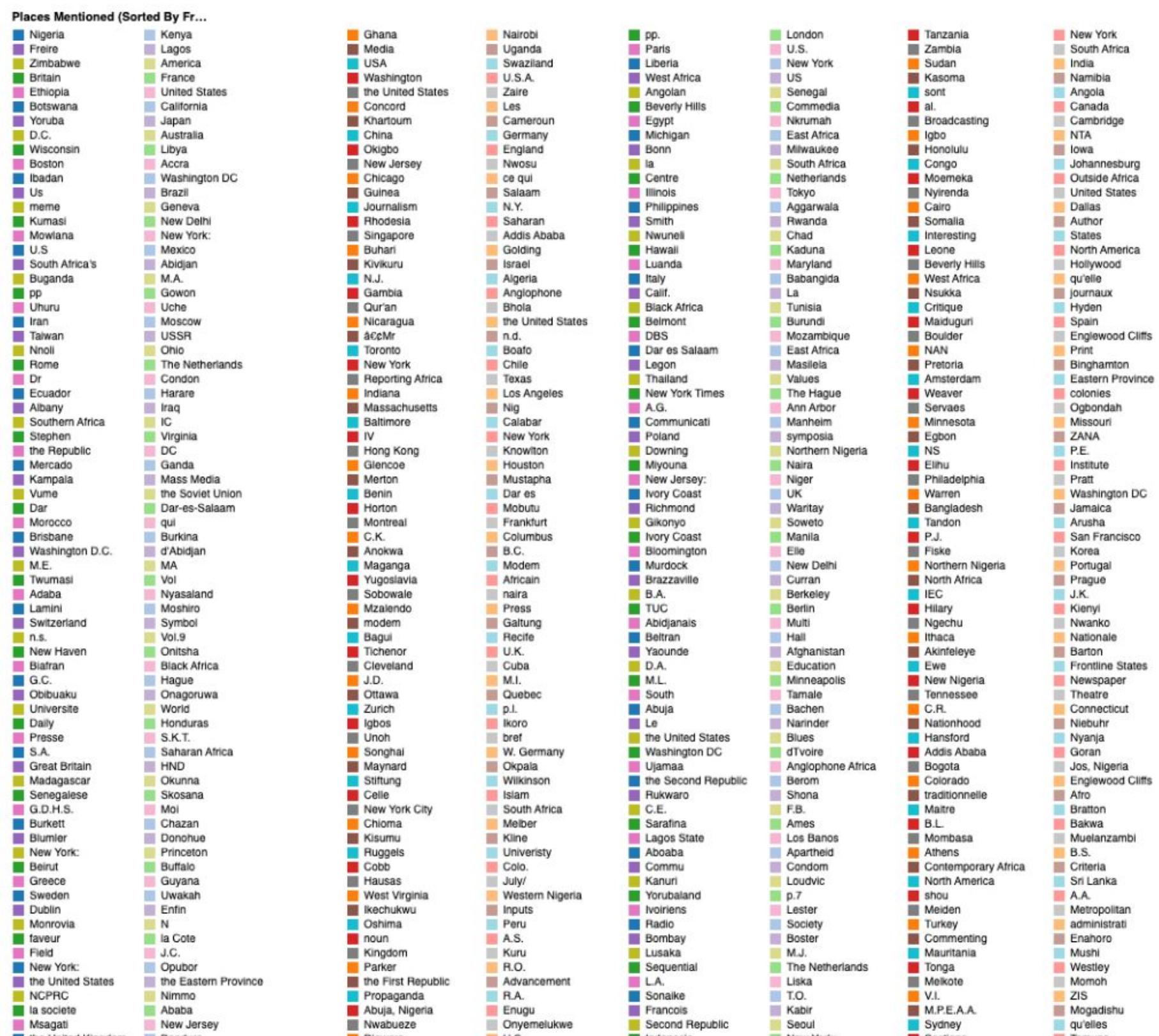
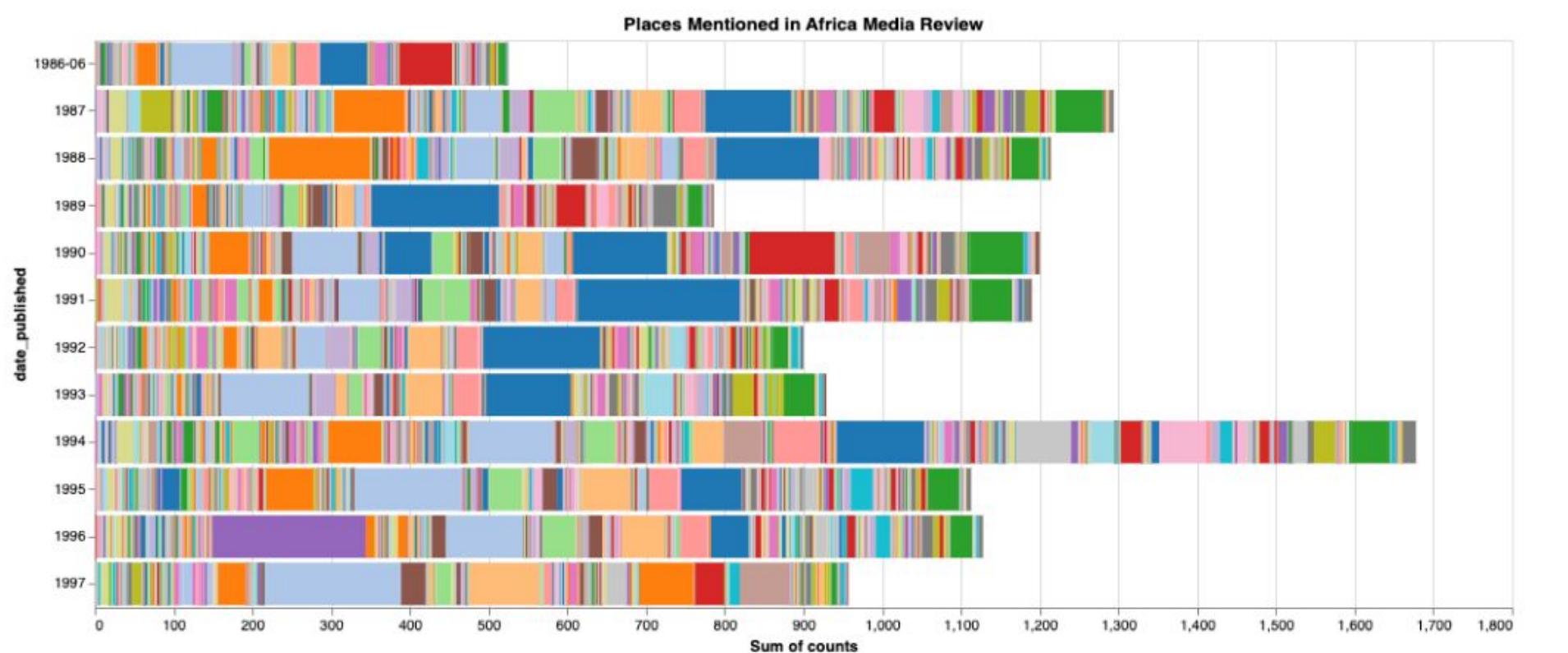
2. OCR Text Length Over Time: Gauging Publishing Activity

A first point of entry is the bar chart displaying the **OCR text length of Africa Media Review articles by publication date (1986–1997)**. In this visualization, each bar corresponds to a particular issue, with the x-axis representing total OCR word count (or token count). Several patterns stand out:

- Initial Years (1986–1988):** Issues in the mid-1980s, shortly after the journal's inception, show relatively high textual volume in some years (notably 1988) and more modest output in others. This fluctuation may reflect the early establishment of editorial procedures, shifting funding opportunities, and the ACCE's ability to commission longer or more numerous articles.
- Variability in the 1990s:** As the 1990s progress, there are spikes (e.g., around 1994–1995) when the total text length surges toward the higher end of the scale. These spikes could indicate special issues or increased interest in debates around the liberalization of African media in the post–Cold War landscape, dovetailing with **Toyin Falola** and **Emmanuel Akyeampong**'s emphasis on the changing political economy of African states.
- Steady Output Despite Political Upheavals:** The overall continuity of publishing, even amid the early 1990s political transitions in many African countries, speaks to the resilience of the editorial team and the consistent scholarly interest in African media. Scholars like **Kwame Karikari** underscore how press freedom debates surged during periods of democratization, likely prompting longer or more in-depth articles.

By quantifying the text length, we gain a proxy measure of publishing volume and editorial intensity—key indicators of the journal's evolving place within Africa-focused scholarship on communication.

Final Results



3. Place Names and Spatial Coverage: NER Insights

A second visualization—a stacked bar chart showing named-entity recognition (NER) results for place names—reveals which locations were most frequently mentioned in *Africa Media Review*. Each row corresponds to a publication date, and the x-axis represents the total number of place-name references extracted from that year's issue(s). Several interpretive points emerge:

- Prominence of African Countries:** Nigeria, Kenya, Ghana, and Zimbabwe appear frequently, reflecting both the intellectual hubs of African communication research and the specific case studies that authors chose. This resonates with **Frederick Cooper's** (2002) observations about the importance of particular national contexts in African historiography—Nigeria, for instance, often exemplifies a large, diverse media ecosystem.
- Transnational Focus:** References to the United States, France, and Britain also feature in the chart—unsurprising given the ongoing influence of former colonial powers, as well as the broader North-South debates about media flows (e.g., NWICO). The presence of places like "Paris," "London," and "Washington" underscores how *Africa Media Review* engaged global power centers, aligning with the postcolonial critique of external media hegemony advanced by **Walter Rodney** (1972) and others.
- Surprising Mentions:** The mosaic includes smaller counts of less-expected place names—possibly one-off references to academic conferences or lesser-known case studies. Their presence reflects the journal's eclectic coverage and openness to varied African and non-African contexts.

Crucially, these data illustrate the journal's **pan-African orientation**: while certain countries dominate, a wide spread of locales attests to diverse contributions. This geographical breadth exemplifies the pan-continental identity of the ACCE and the journal's editorial ethos of inclusivity.

5. Incorporating the Visualizations into the Broader Historiography

5.1. Linking Publishing Patterns to Decolonization Debates

The variations in OCR text length hint at shifting editorial capacities, possibly influenced by external factors like political instability or academic funding cycles. As **Bethwell A. Ogot** (1963) and **Jean Suret-Canale** (1964) have shown, the production of African history (and by extension, African media scholarship) often reflects wider currents of state policy, donor interests, and intellectual politics. The bar chart thus offers an empirical measure to anchor broader claims about the challenges of sustaining African-led journals in an uneven global publishing ecosystem.

5.2. Geographic Breadth as a Pan-African Imperative

NER-based insights reinforce how *Africa Media Review* contributed to a pan-continental discourse, repeatedly featuring examples across Anglophone, Francophone, and Lusophone contexts. This approach parallels **Ngũgĩ wa Thiong'o's** (1986) call to transcend linguistic and colonial divides in forging a unified African intellectual space. The data reveal that *Africa Media Review* not only critiqued media imperialism but also showcased empirical realities from diverse African settings—an embodiment of the journal's commitment to decolonizing media scholarship.

5.3. Thematic Clustering and the Evolution of Communication Research

Finally, the UMAP scatter plot underscores the journal's role in charting—and sometimes anticipating—major shifts in communication research. As Africa entered the 1990s, transitions to multiparty democracy, liberalized economies, and the early adoption of digital communication likely spurred new thematic clusters in the journal. Such continuity and change echo **Emmanuel Akyeampong's** (2000) reminder that African academic discourse is neither static nor parochial; it adapts to political upheavals, technological advancements, and scholarly trends.

Evaluating Our Experiments with AI

The Pudding

Can an AI make a data-driven, visual story?

[Read Claude's story >](#)

By [Russell Samora](#) and [Michelle Pera-McGhee](#)

With LLMs (henceforth, **AI**) in the spotlight – blowing minds, raking in venture capital, and prompting existential crises – we at The Pudding were curious: Will this make our jobs obsolete? How scared should we be that someone else could replicate what we do without the time, training, and expertise we have?

Basically, **can an AI make a data-driven, visual story, much like we do at The Pudding?** What does it actually “do,” and how well does it do it?

So, we tried replacing ourselves with [Claude](#), an AI from Anthropic.

Institutional Policies

- **Data Compliance**
 - State/federal laws & institutional regulations
 - Acceptable use & institutional data policies
 - Use of third-party commercial gen AI
 - Governance, risk, and compliance
 - Ethical considerations align with university's values
 - Data classifications
 - Public data, confidential/private, research data
 - **Risks and Limitations**
 - Recommended practices
 - Consider impact, critically evaluate info obtained from AI, disclosure requirements for using gen AI in academic work, adhere to updated guidance
 - **Preparing Publications**
 - Journal policies on AI use
 - Authorship and originality
 - Data integrity verification
 - Disclosure of AI involvement
 - Review for biases or inaccuracies
 - **Equitable, Inclusivity, and Ethical Use**
 - Lack of transparency
 - Bias (and hallucinations)
 - Ethical sourcing of training material
- <https://research.msu.edu/generative-ai/guidance>

Broader Considerations

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?



Emily M. Bender^{1*}, Timnit Gebru^{2*},
Angelina McMillan-Major¹, Shmargaret Shmitchell³

¹ University of Washington ² Black in AI ³ The Aether
*These authors contributed equally.

Issue 3

PARROTS

Are Large Language Models Our Limit Case?

Lauren Klein

August 2022

[doi:10.5281/zenodo.6567985](https://doi.org/10.5281/zenodo.6567985)

[TXT](#) [PDF](#)

How is it possible to go forward with the development and use of large language models, given the clear evidence of just how biased, how incomplete, and how harmful — to both people and the planet — these models truly are? This is the question that “Stochastic Parrots” productively sets in motion, and that has continued to reverberate throughout the artificial intelligence (AI) and machine learning (ML) community since the paper’s release.¹

This Isn't a Calculator: AI as a Cultural Technology

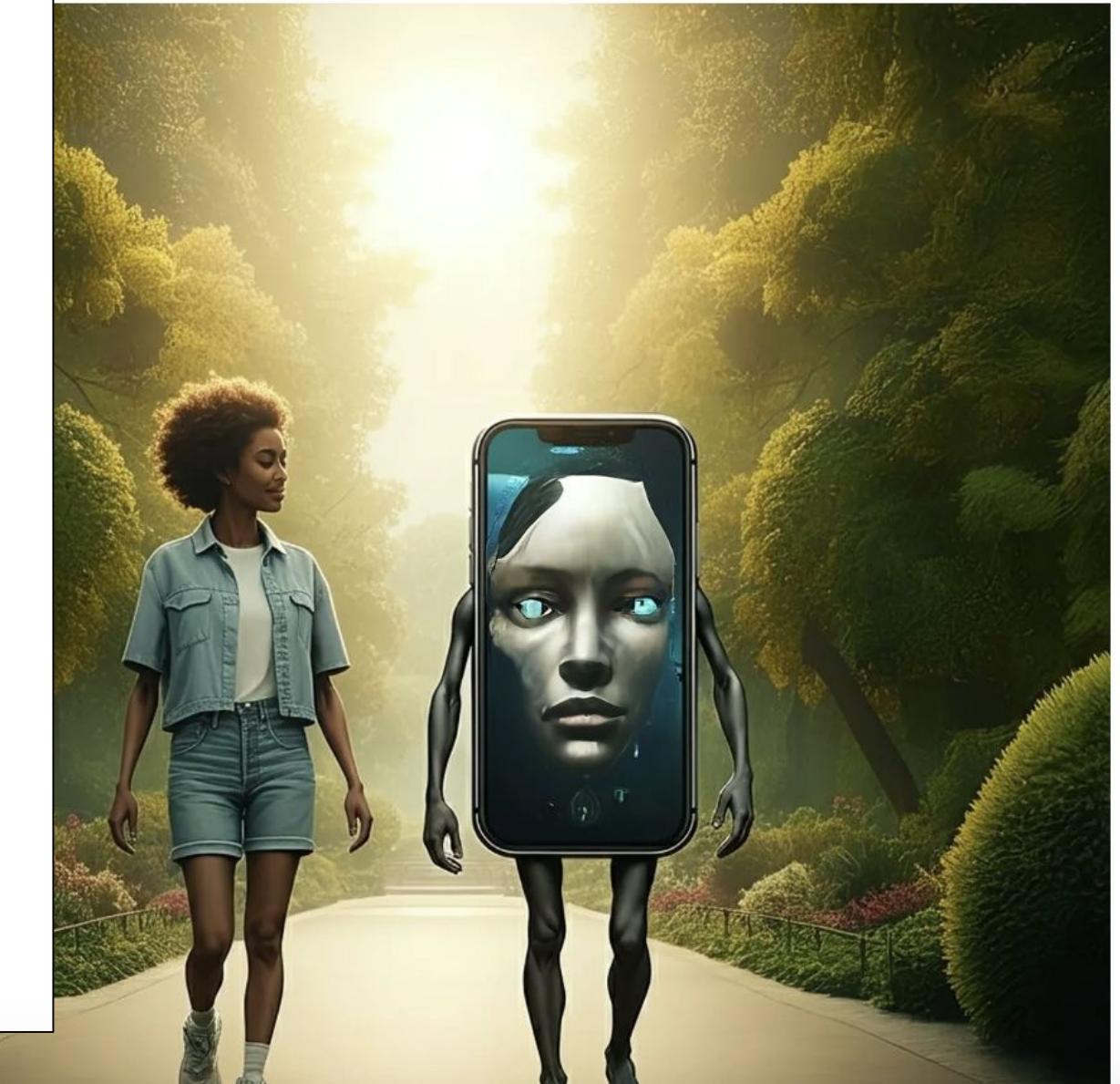
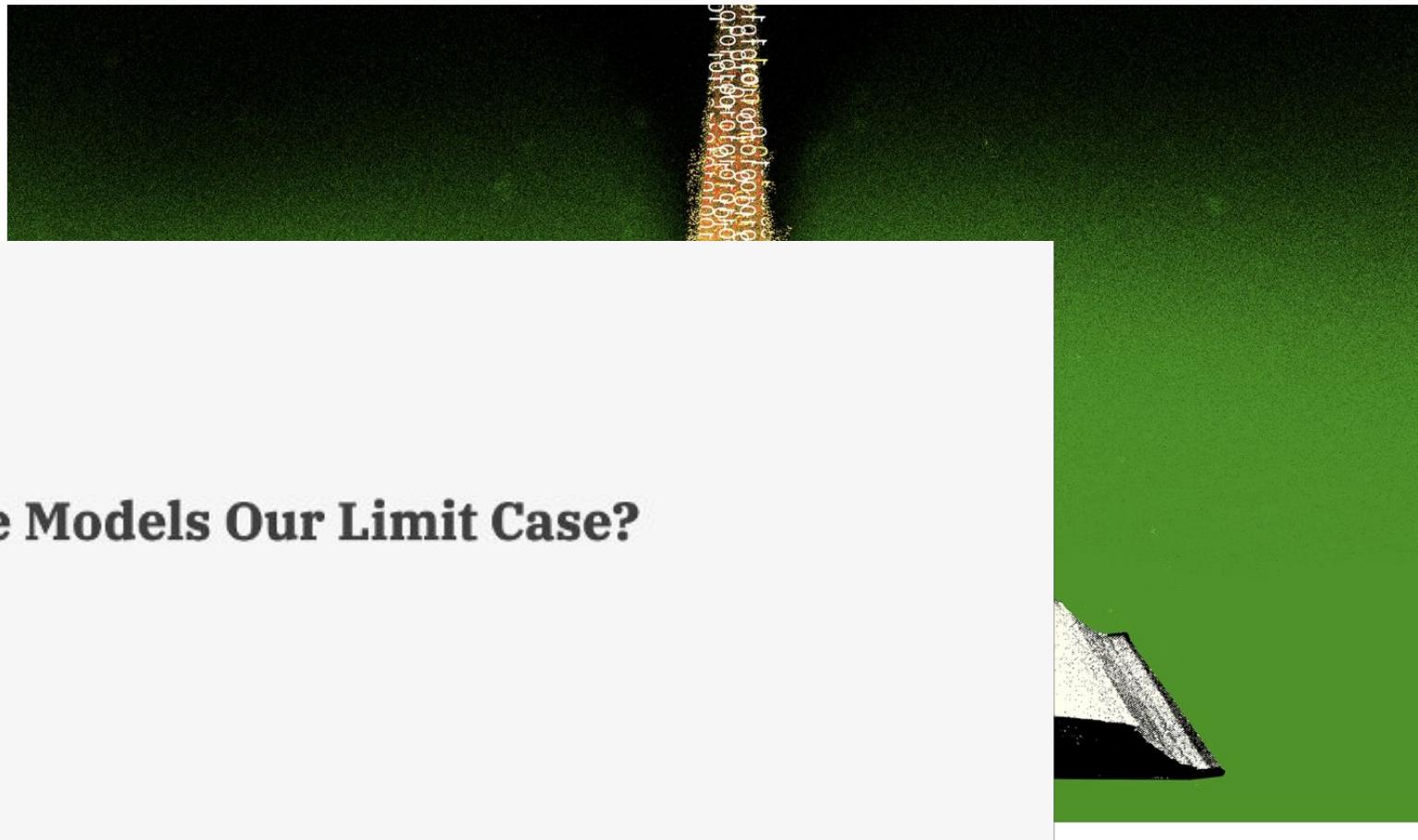


MARC WATKINS
SEP 29, 2024

Prepare for the Textpocalypse

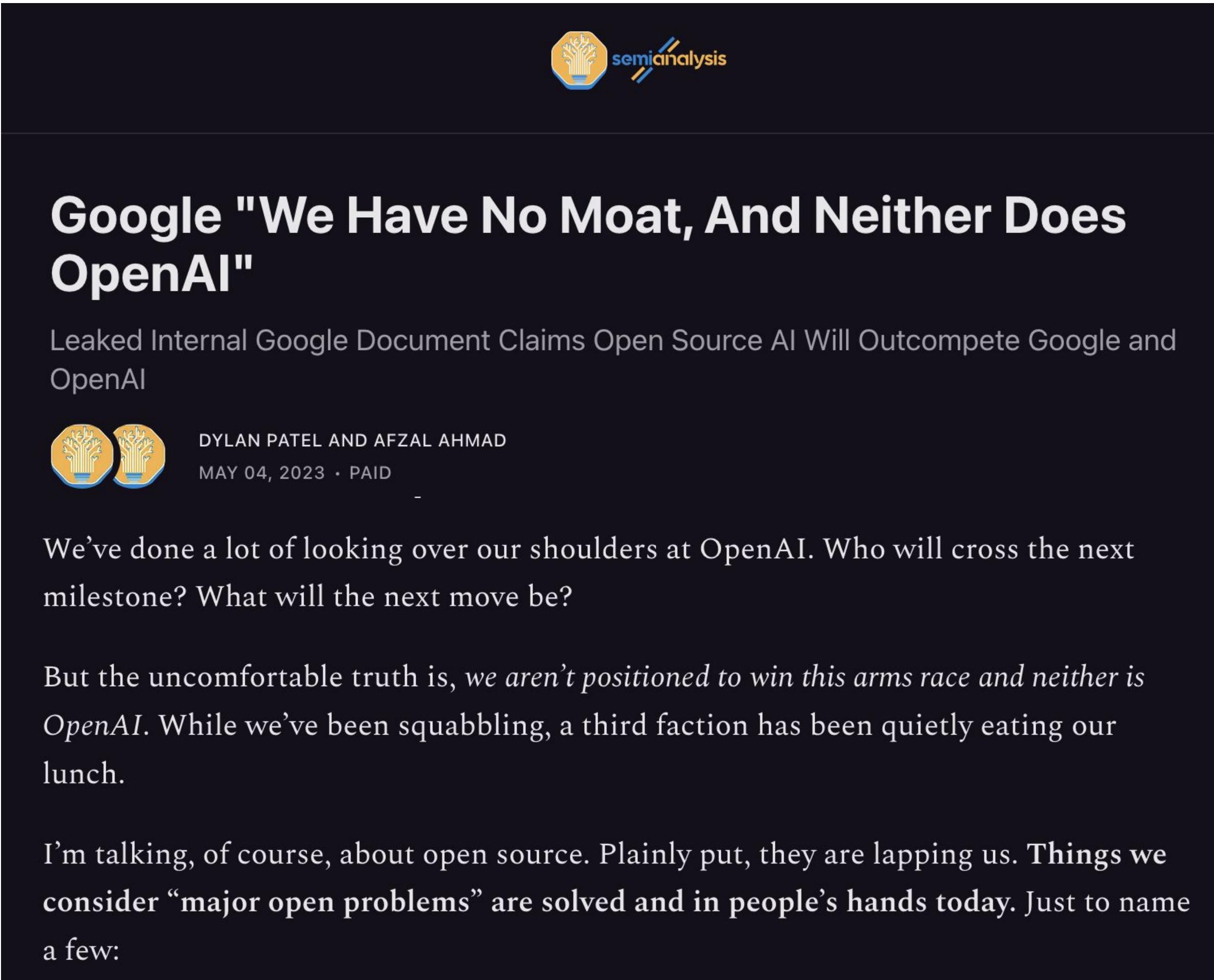
Our relationship to writing is about to change forever; it may not end well.

By Matthew Kirschenbaum

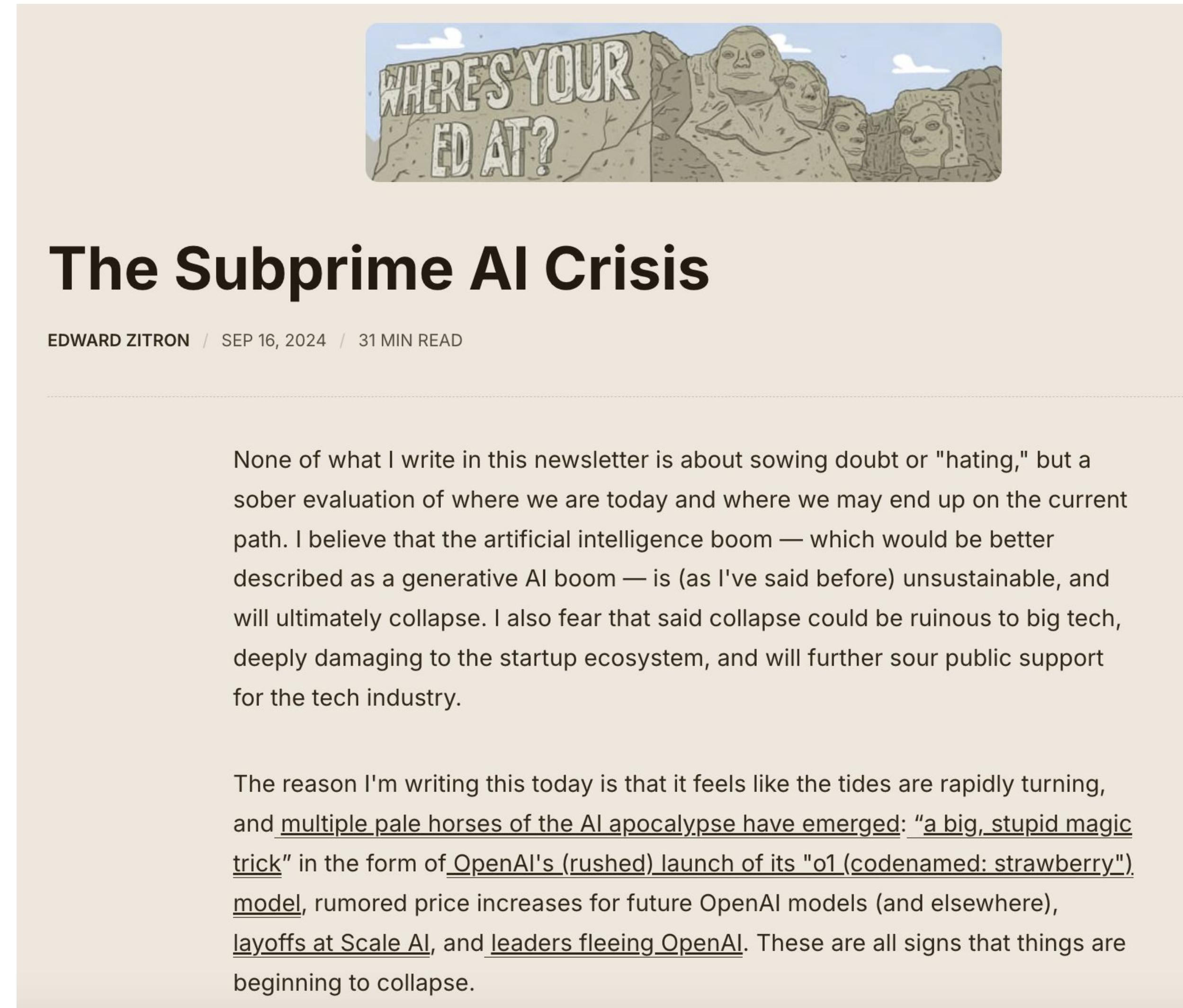


Share

Broader Considerations



The screenshot shows a dark-themed news article. At the top left is the semianalysis logo, which consists of two stylized yellow tree icons and the word "semianalysis". The main title is "Google 'We Have No Moat, And Neither Does OpenAI'" in large white font. Below the title is a subtitle: "Leaked Internal Google Document Claims Open Source AI Will Outcompete Google and OpenAI". The authors are listed as "DYLAN PATEL AND AFZAL AHMAD" with a small photo of them next to their names. The publication date is "MAY 04, 2023 · PAID". The article begins with a quote: "We've done a lot of looking over our shoulders at OpenAI. Who will cross the next milestone? What will the next move be?" followed by a paragraph about OpenAI's competition with Google.



The illustration features a light brown background with a blue sky and clouds. On the left, there's a large, textured rock formation with the words "WHERE'S YOUR ED AT?" carved into it in a bold, blocky font. To the right of the text, the iconic faces of George Washington, Thomas Jefferson, Abraham Lincoln, and Theodore Roosevelt are carved into the side of the mountain. The overall tone is satirical, referencing the famous "Where's Waldo?" book series.

The Subprime AI Crisis

EDWARD ZITRON / SEP 16, 2024 / 31 MIN READ

None of what I write in this newsletter is about sowing doubt or "hating," but a sober evaluation of where we are today and where we may end up on the current path. I believe that the artificial intelligence boom — which would be better described as a generative AI boom — is (as I've said before) unsustainable, and will ultimately collapse. I also fear that said collapse could be ruinous to big tech, deeply damaging to the startup ecosystem, and will further sour public support for the tech industry.

The reason I'm writing this today is that it feels like the tides are rapidly turning, and multiple pale horses of the AI apocalypse have emerged: "a big, stupid magic trick" in the form of OpenAI's (rushed) launch of its "o1 (codenamed: strawberry") model, rumored price increases for future OpenAI models (and elsewhere), layoffs at Scale AI, and leaders fleeing OpenAI. These are all signs that things are beginning to collapse.

***Discuss more in final
session***

Thanks for participating!