Telling Stories with Data

In an AI Age

Rohan Alexander, University of Toronto

eCOTS, 16 June 2026

If the first line of your R script is

```
setwd("C:\Users\jenny\path\that\only\I\have")
```

I* will come into your office and
SET YOUR COMPUTER ON FIRE 🔥.

* or maybe Timothée Poisot will

Where were we a year ago?

“A lot of jobs will go away”

Sam Altman, 2025

“I'm delighted to be wrong about this.”

Sam Altman, 2026

https://www.reuters.com/world/asia-pacific/openais-altman-says-ai-unlikely-lead-jobs-apocalypse-2026-05-26/

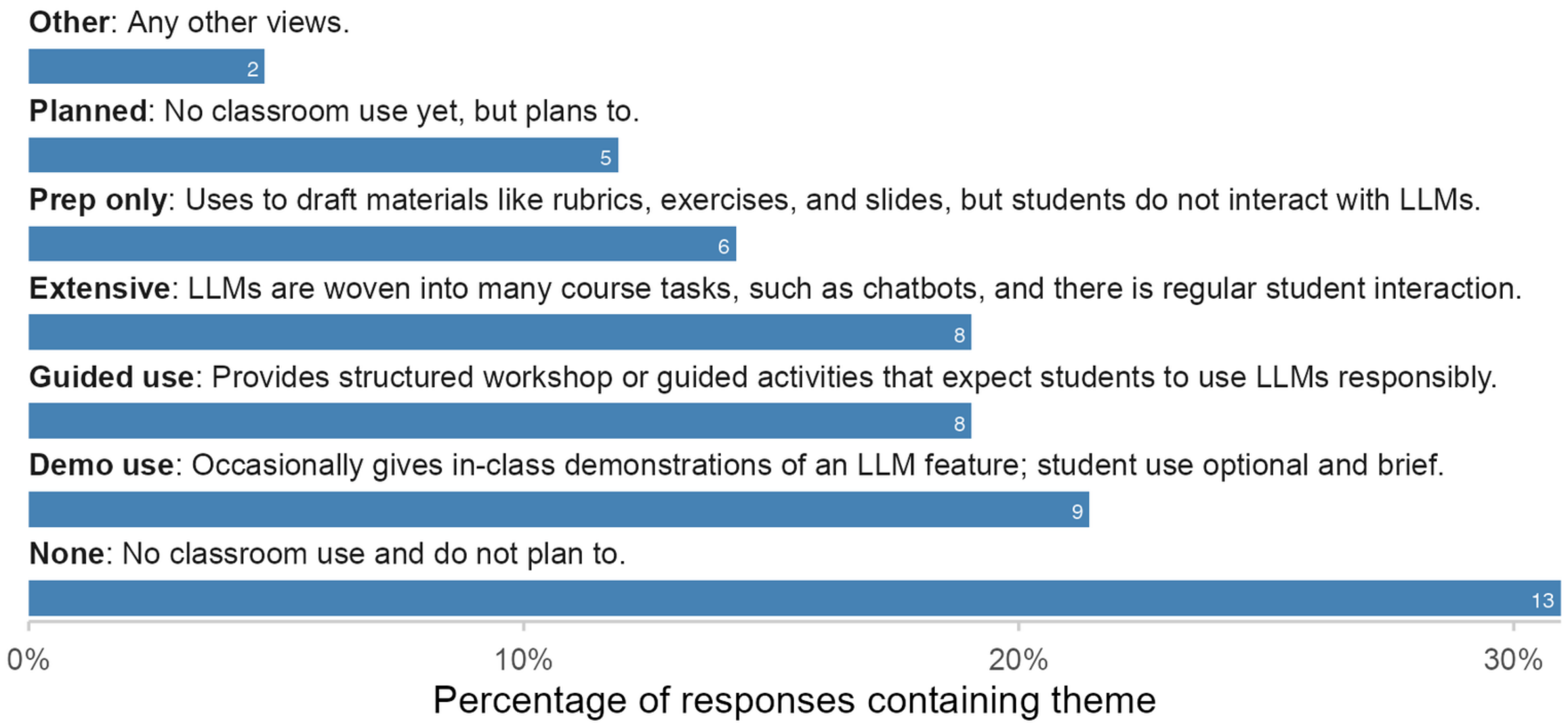
Prompting the Professoriate

A Qualitative Study of Instructor Perspectives on LLMs in Data Science Education

- 42 instructors, 33 institutions, 9 countries, June and July 2025.
- LLMs as inevitable classroom tools—comparable to calculators.
- Others worry about de-skilling, misplaced confidence, and uneven integration across institutions.
- Around 58% have already introduced demonstrations, guided activities, or made extensive use of LLMs in their courses.
- 31% have not used LLMs to teach students and do not plan to.

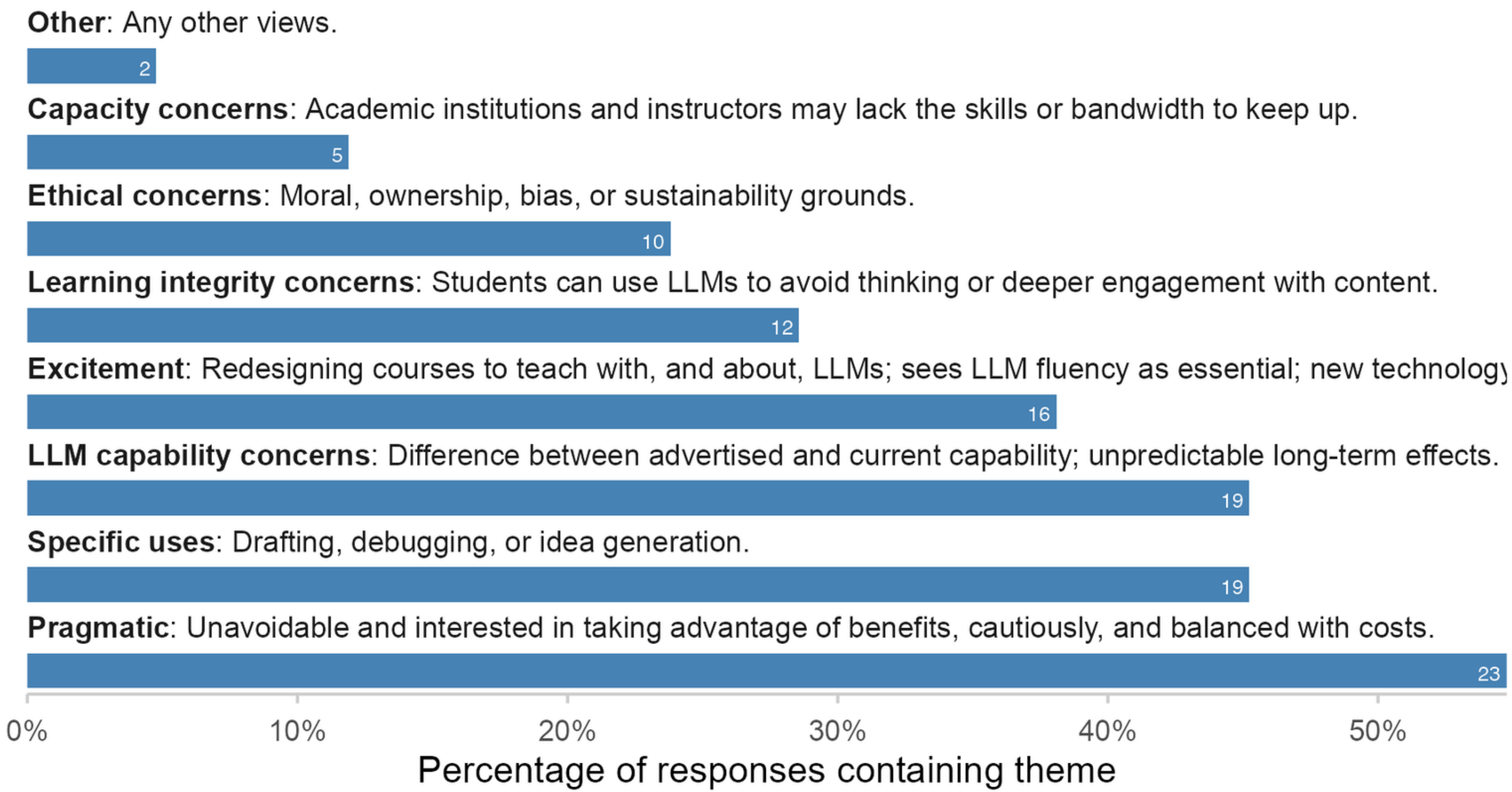
https://doi.org/10.1162/99608f92.c8780a45

Have you used LLMs to teach students, if so, how have you used LLMs to teach students?

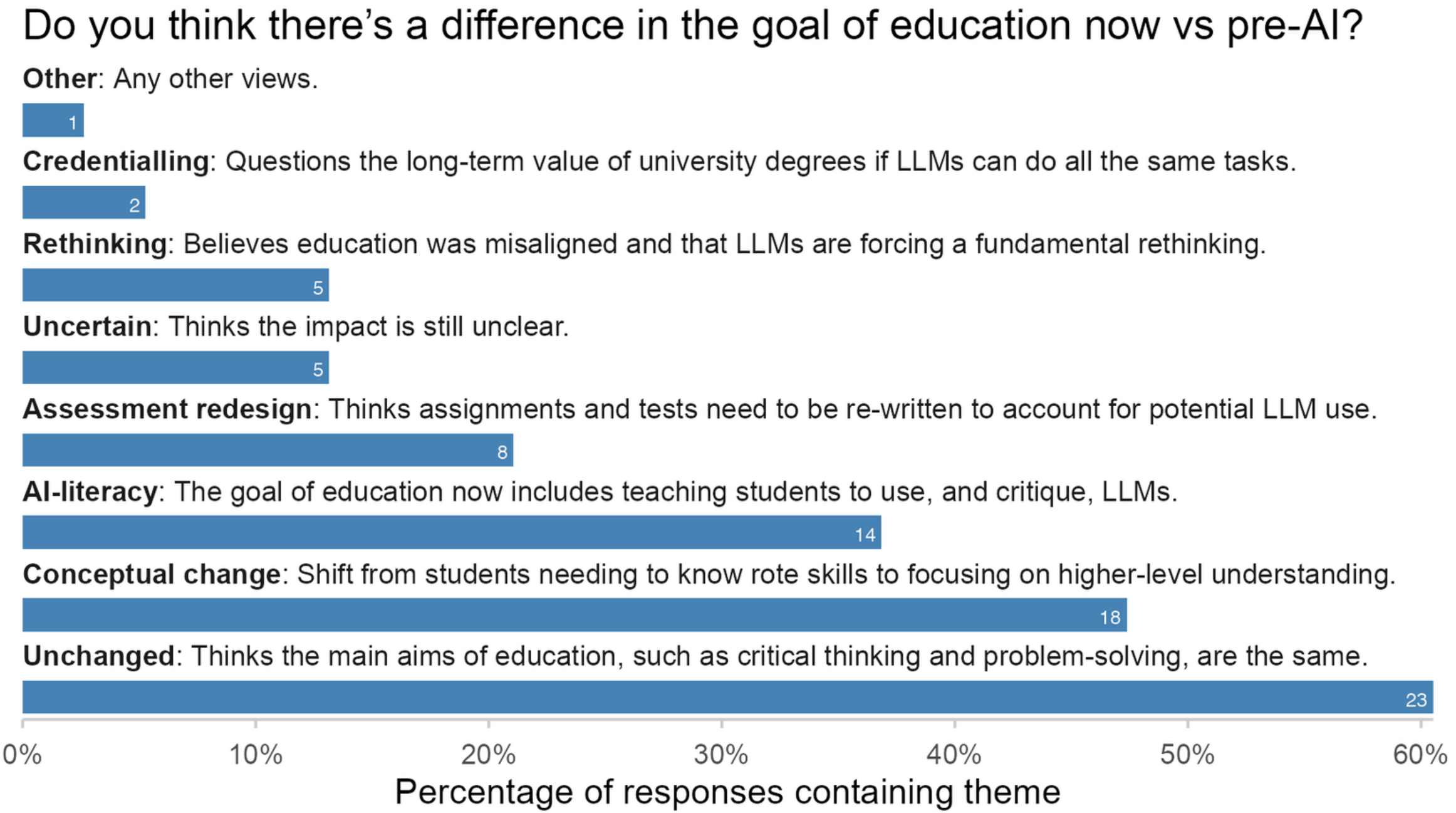


https://doi.org/10.1162/99608f92.c8780a45

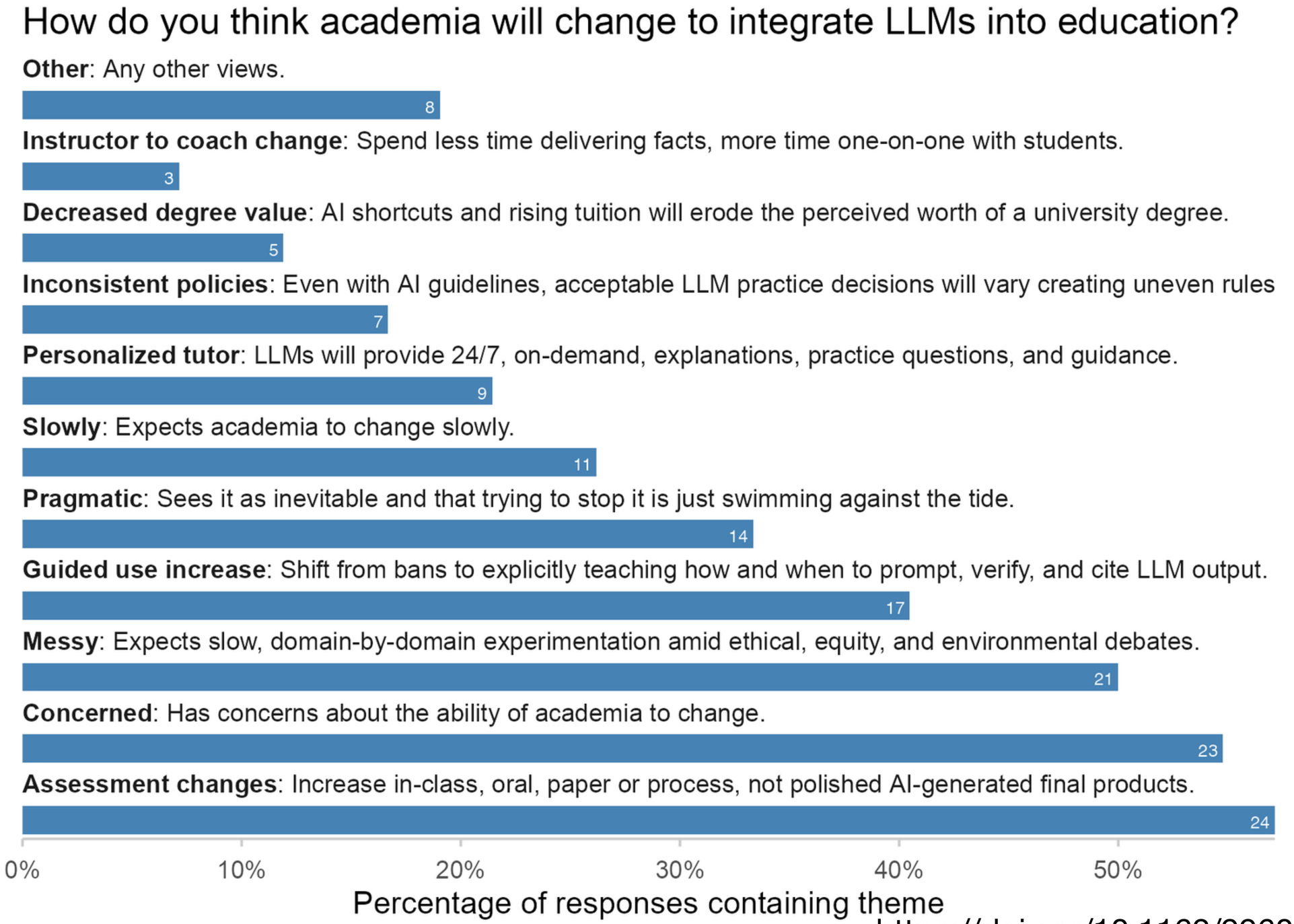
What are your general views on LLMs?



https://doi.org/10.1162/99608f92.c8780a45



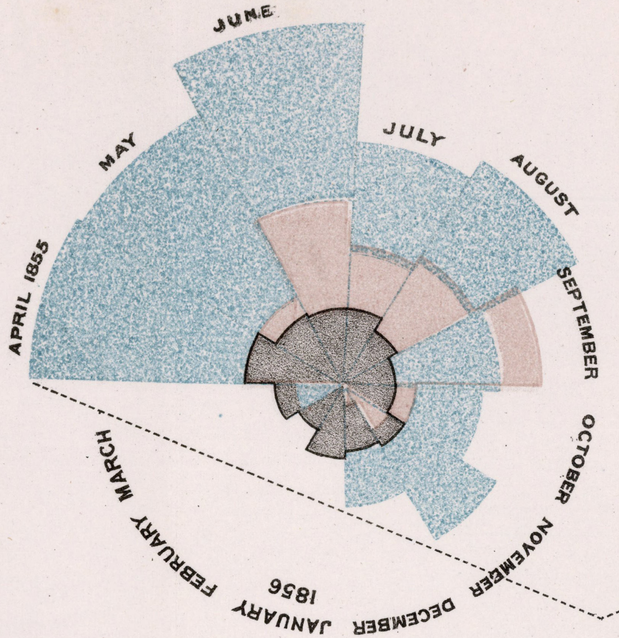
https://doi.org/10.1162/99608f92.c8780a45



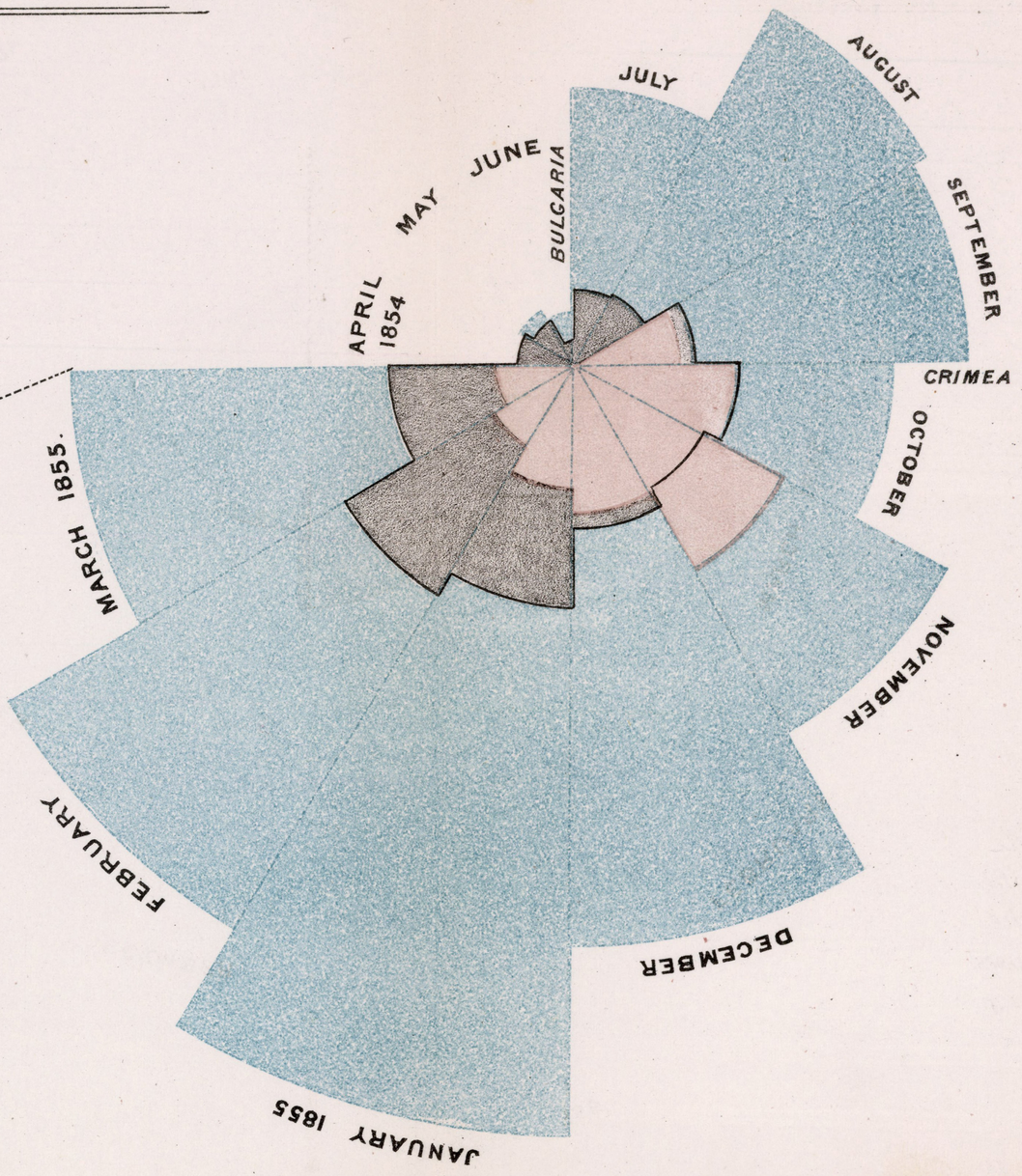
https://doi.org/10.1162/99608f92.c8780a45

DIAGRAM of the CAUSES of MORTALITY IN THE ARMY IN THE EAST.

2.
APRIL 1855 to MARCH 1856.



1.
APRIL 1854 to MARCH 1855.



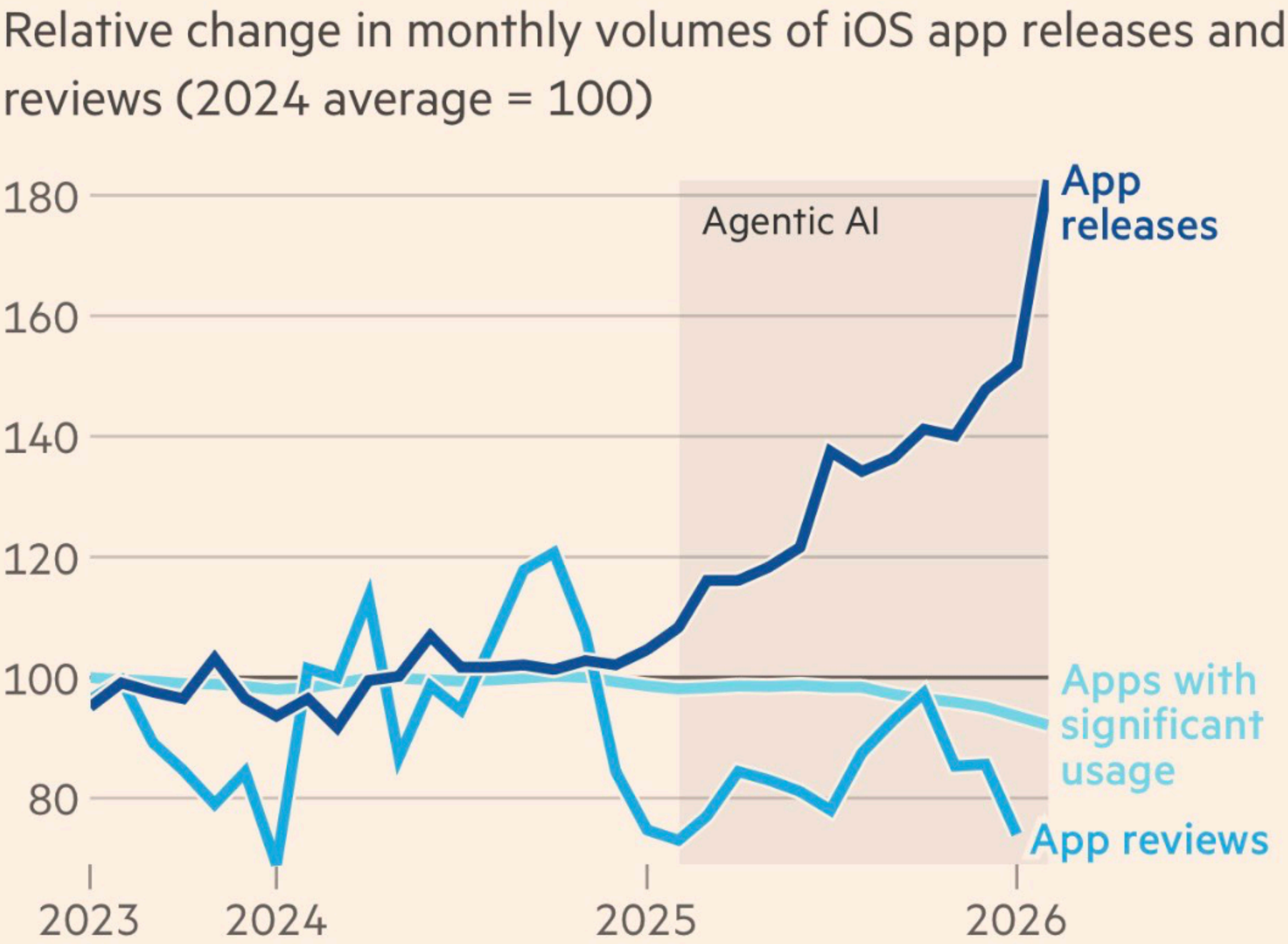
The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds; & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red; in January & February 1855, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.



[12] https://www.nber.org/papers/w35275 and Financial Times

The GenAI Divide

STATE OF AI IN BUSINESS 2025

MIT NANDA

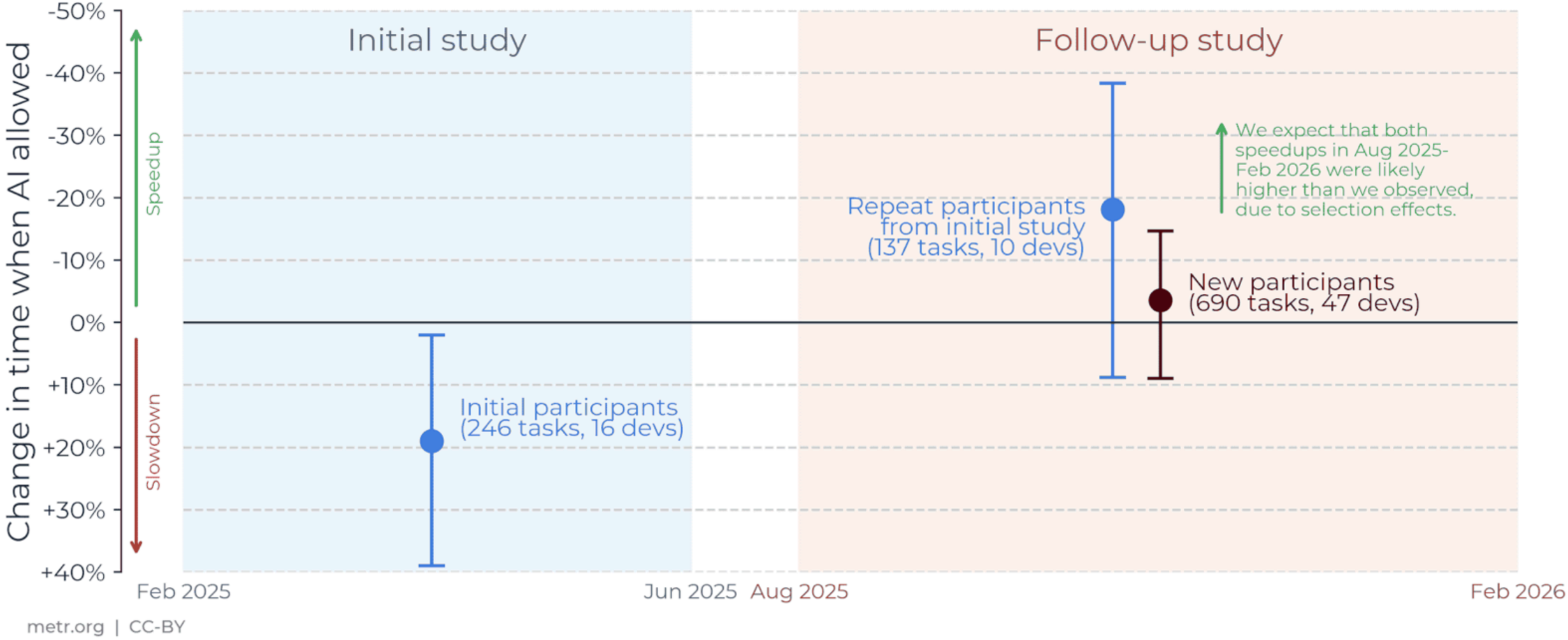
Aditya Challapally
Chris Pease
Ramesh Raskar
Pradyumna Chari
July 2025

https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf

Late-2025 AI likely accelerates open-source developers, but selection effects make our follow-up results unreliable

METR

Developers in our follow-up RCT increasingly declined to work without AI, mostly due to anticipated productivity loss and to a lesser extent reduced pay. The resulting selection effect likely leads our new results to underestimate speedup.



metr.org | CC-BY

So can we trust a single answer?

Same Prompt, Different Outcomes: Evaluating the Reproducibility of Data Analysis by LLMs

Question: LLMs can do many aspects of data science, but is their analysis reproducible?

Method:

1. We develop an evaluation task.
2. We use the many-analysts paradigm to evaluate how three LLMs approach an identical data science task..

Answer: Substantial variation in how different LLMs approach data cleaning, variable definition, and statistical analysis, even with identical instructions.

Background
Many-analysts paradigm

- Normal social science:
  - Different scholars, evaluate different datasets.
  - Find differences! 😊
- Many-analysts studies:
  - Different scholars, evaluate the same dataset.
  - Still find differences! 😢

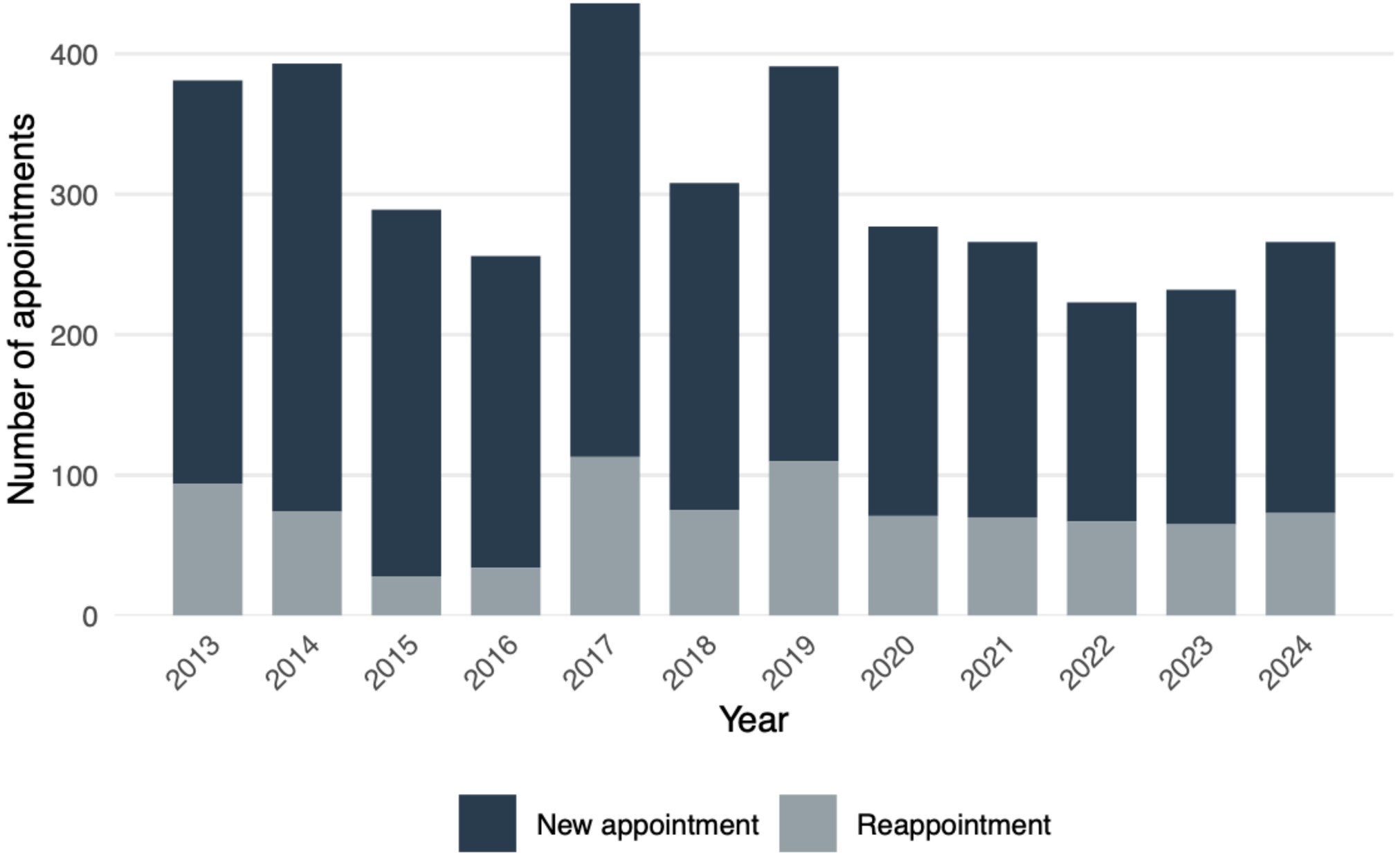
Background
Silberzahn et al. (2018)

- Set-up:
  - Do soccer referees give more red cards to dark skinned players?
  - 29 research teams were given identical data.
- Results:
  - Effect sizes (odds ratios) ranging from 0.89 to 2.93.
  - 20 teams found a statistically significant effect.
  - 9 teams found no significant relationship.

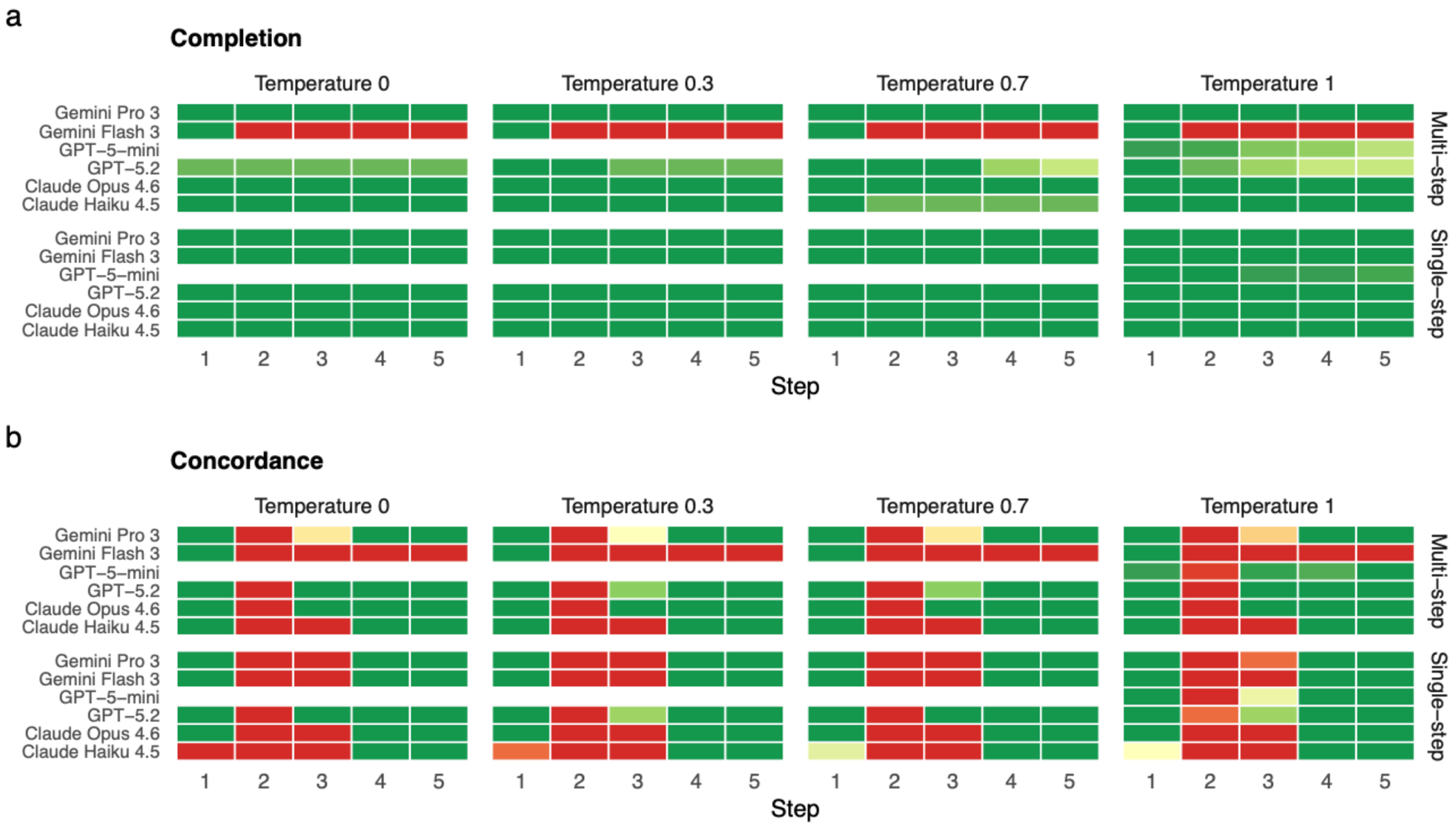
Our set-up

- Which government branch in New Brunswick most frequently reappoints past appointees, and is this trend increasing or declining?
- To answer this an LLM must do:
  - data integration,
  - variable construction, and
  - statistical analysis.

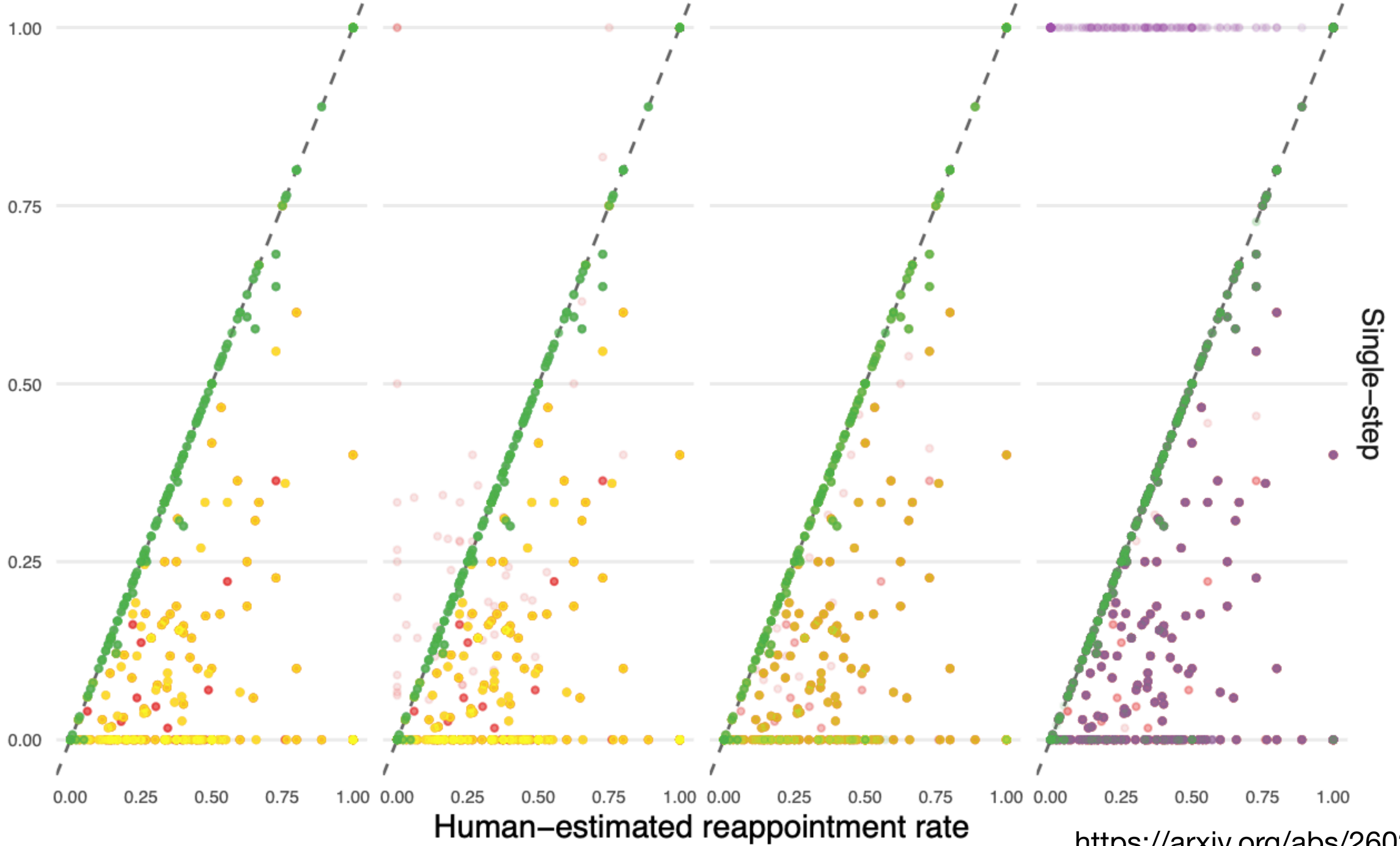
https://arxiv.org/abs/2602.14349



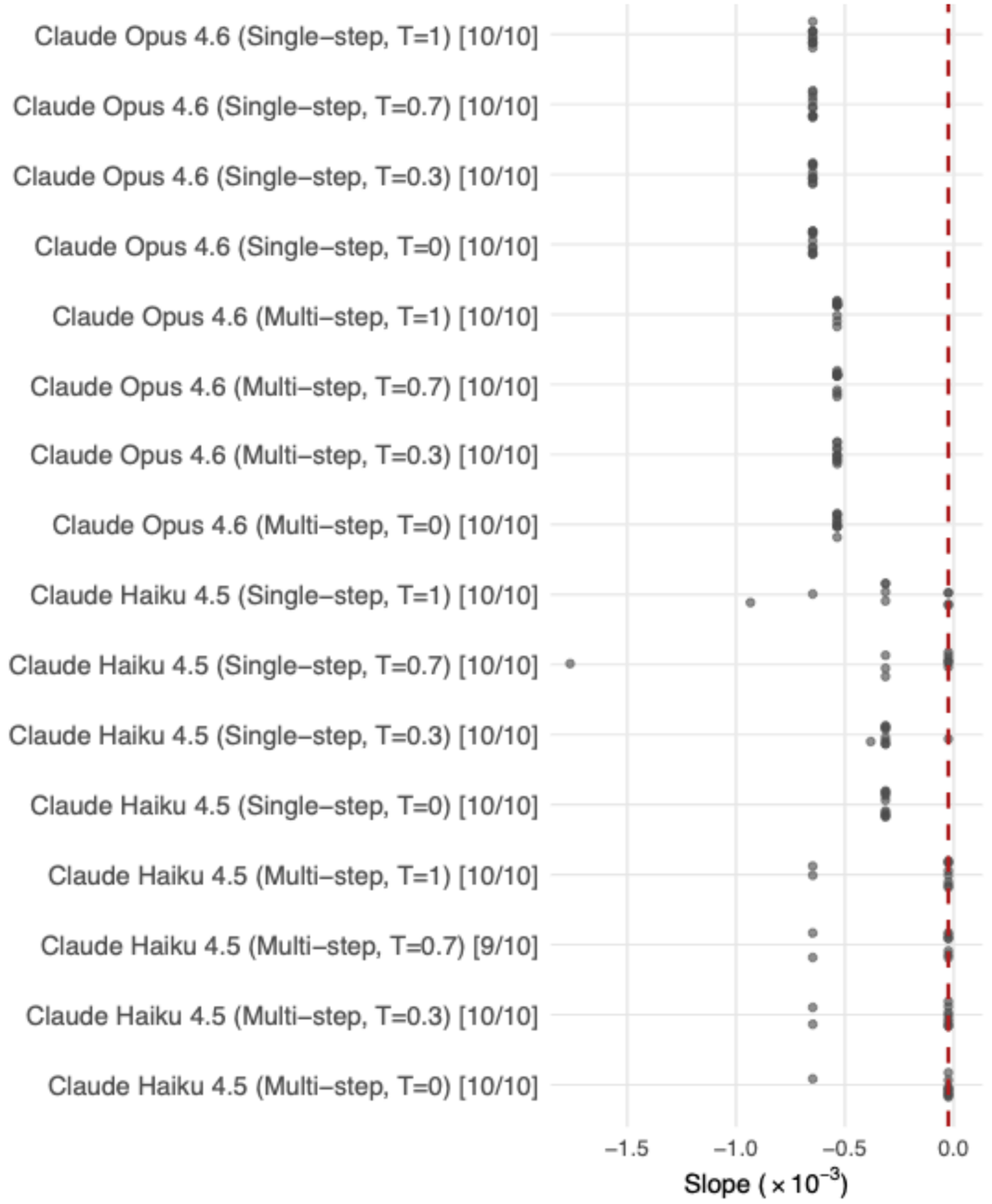
https://arxiv.org/abs/2602.14349



https://arxiv.org/abs/2602.14349



https://arxiv.org/abs/2602.14349

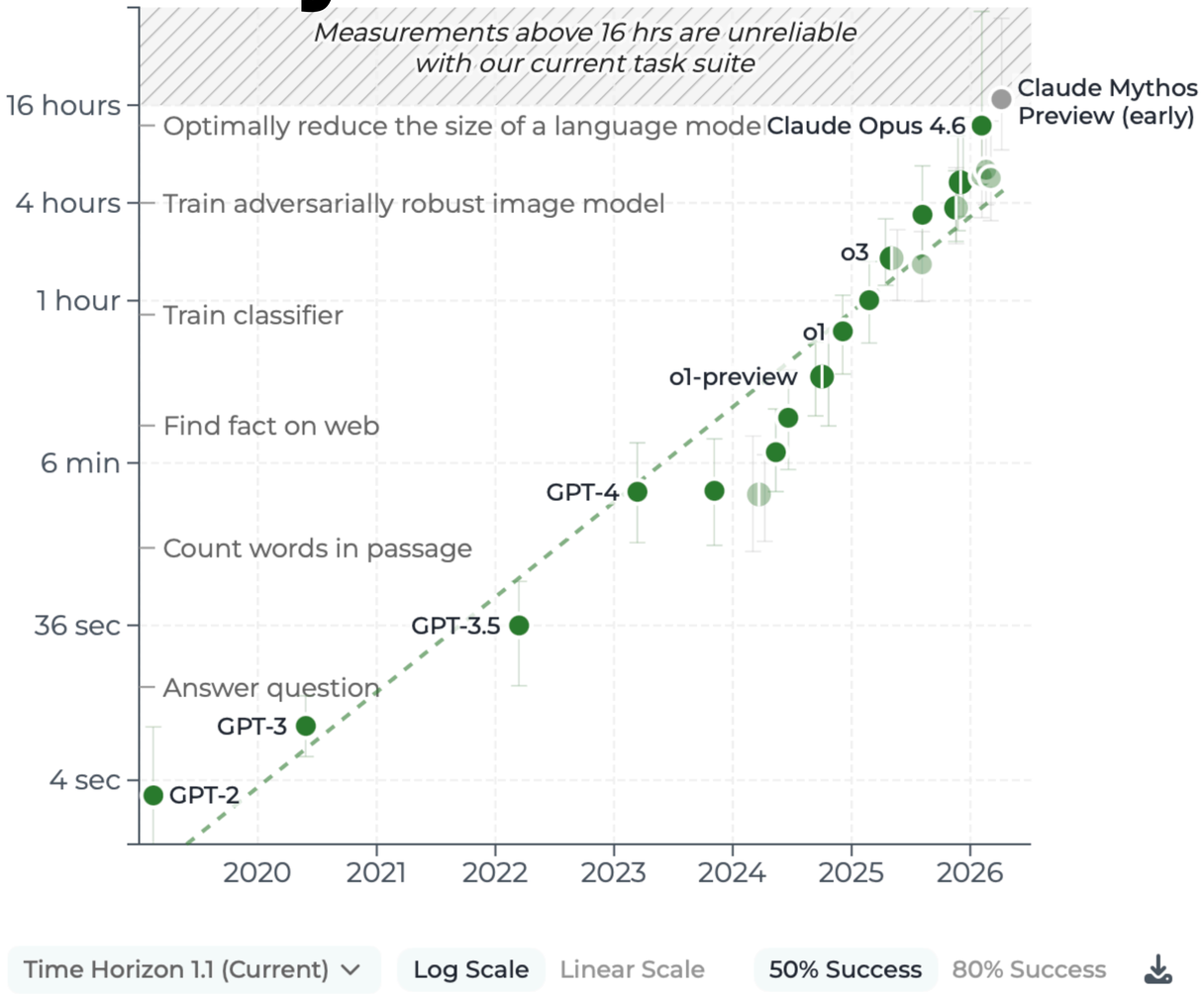


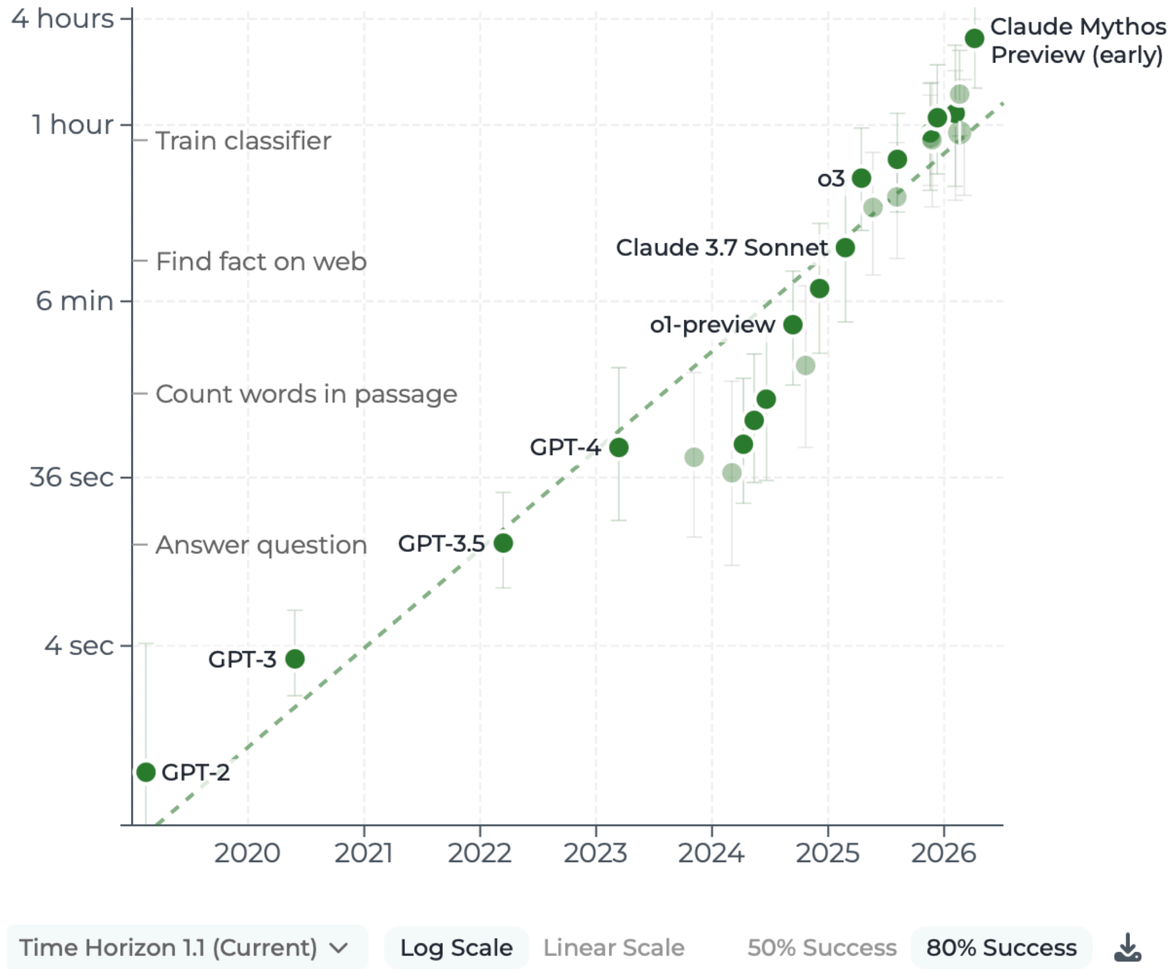
https://arxiv.org/abs/2602.14349

Key takeaways

- Output differs even between executions with the same settings, and that the results can be different to that of a human.
- Our results demonstrate the need to run many executions of the same task when conducting data analysis with LLMs.
- Different models performed differently, and so our results also emphasize the importance of running any data analysis with models from at least two different providers.
- That all said, the estimates from the LLMs tended to differ in particular ways.

How you evaluate matters





https://arxiv.org/abs/2503.14499

Benchmarking AI Performance on End-To-End Data Science Projects

Question: Can generative AI models generate end-to-end data science projects?

Method:

1. Create a benchmark of 40 end-to-end marked data science projects.
2. Develop an automated grading pipeline.
3. Systematically evaluate six generative AI models on the quality of the data science projects.

Answer: It varies considerably by model. Most recent models did well on structured tasks, but there were considerable differences on tasks that needed judgment.

https://arxiv.org/abs/2602.14284

Approach

1. Benchmark construction
2. Auto-grader development and validation
3. Project generation pipeline

https://arxiv.org/abs/2602.14284

Benchmark construction
Based on student projects

- To construct our benchmark we collect 40 projects completed by university students, anonymize them, and then develop a 17-category rubric and two human experts use that to mark each project and then resolve any differences.
- The task is to complete a data analysis project using data from Open Data Toronto.
- This task is adapted from an upper-year undergraduate assignment at the University of Toronto.

https://arxiv.org/abs/2602.14284

Benchmark construction
Expectations

- Each project must include Python scripts to:
  - Download a dataset from Open Data Toronto.
  - Clean and prepare the data.
  - Test the dataset for obvious issues.
- The paper itself is written as a Quarto document which compiles to a PDF.

Group	Category	Max points
Reproducibility	Reproducible workflow	5
	Tests	3
	Author, date, and repo	1
Presentation	Title	2
	Abstract	4
	Introduction	4
Analysis	Measurement	2
	Data	4
	Discussion	4
Graphs	Data tables	2
	Results figures	4
	Captions	2
Referencing	Cross-references	1
	Bibliography	2
	Python cited	1
Writing	Professionalism	3
	Banned words	1

Auto-grader development and validation
Calibration

- Assessing the projects generated by the generative AI models in a consistent, scalable way required developing and validating an auto-grader.
- We use LLM-as-a-judge with per-category evaluation.
- Our benchmark calibrates the auto-grader.
- This resulted in an automated grading pipeline.

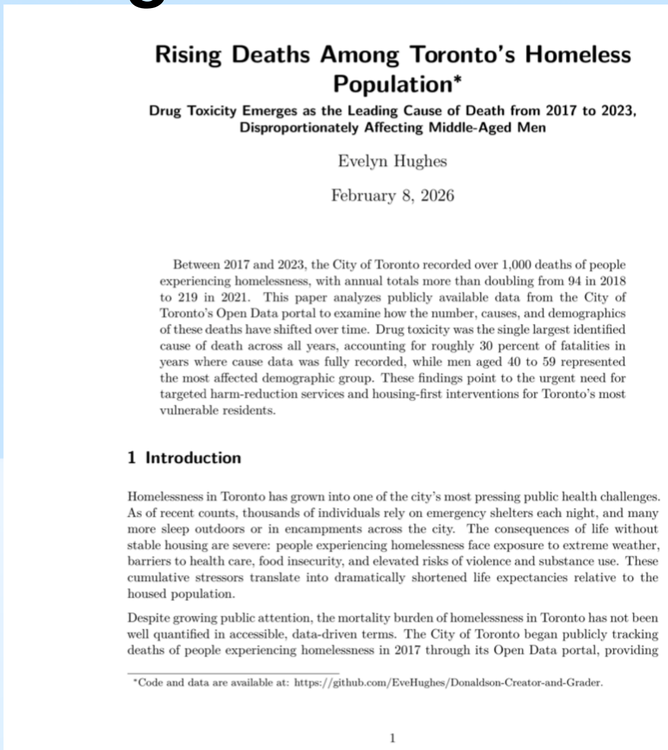
Project generation pipeline

Each step is a separate API call

Phase	Step	Output
Data acquisition	1. Planning	Dataset selection, research question
	2. Download	01-download.py
	3. Clean	02-clean.py
Quality assurance	4. Testing	03-test.py (pytest suite)
	5. Requirements	requirements.txt or uv
Documentation	6. References	references.bib
	7. Figures	figures.qmd (visualizations)
	8. Paper	paper.qmd (final manuscript)
	9. README	README.md (reproduction guide)

Example generated papers

The good



Rising Deaths Among Toronto's Homeless Population*

Drug Toxicity Emerges as the Leading Cause of Death from 2017 to 2023, Disproportionately Affecting Middle-Aged Men

Evelyn Hughes

February 8, 2026

Between 2017 and 2023, the City of Toronto recorded over 1,000 deaths of people experiencing homelessness, with annual totals more than doubling from 94 in 2018 to 219 in 2021. This paper analyzes publicly available data from the City of Toronto's Open Data portal to examine how the number, causes, and demographics of these deaths have shifted over time. Drug toxicity was the single largest identified cause of death across all years, accounting for roughly 30 percent of fatalities in years where cause data was fully recorded, while men aged 40 to 59 represented the most affected demographic group. These findings point to the urgent need for targeted harm-reduction services and housing-first interventions for Toronto's most vulnerable residents.

1 Introduction

Homelessness in Toronto has grown into one of the city's most pressing public health challenges. As of recent counts, thousands of individuals rely on emergency shelters each night, and many more sleep outdoors or in encampments across the city. The consequences of life without stable housing are severe: people experiencing homelessness face exposure to extreme weather, barriers to health care, food insecurity, and elevated risks of violence and substance use. These cumulative stressors translate into dramatically shortened life expectancies relative to the housed population.

Despite growing public attention, the mortality burden of homelessness in Toronto has not been well quantified in accessible, data-driven terms. The City of Toronto began publicly tracking deaths of people experiencing homelessness in 2017 through its Open Data portal, providing

*Code and data are available at: https://github.com/EveHughes/Donaldson-Creator-and-Grader.

1

(a) Claude Opus 4.6, Project 5 (87%)

Deaths Among People Experiencing Homelessness in Toronto, 2017–2023*

Deaths rose sharply through 2021, drug toxicity dominated recorded causes, and monthly counts were highest in winter and late spring

Evelyn Hughes

February 8, 2026

Using the City of Toronto's Open Data Toronto dataset on deaths of people experiencing homelessness, I summarize patterns by year (2017–2023), month, gender, age group, and recorded cause of death. The analysis shows a steady rise from 2018 to a peak in 2021, followed by declines in 2022 and 2023 that still leave annual totals above pre-2020 levels. Drug toxicity is the largest recorded cause category in the cause-of-death table, and monthly totals show clustering in winter and late spring, suggesting that risk is not evenly distributed across the calendar. These descriptive results matter because they help orient public health and housing responses toward when (seasonality), who (demographics), and what (recorded causes) are most associated with mortality among people experiencing homelessness in Toronto.

1 Introduction

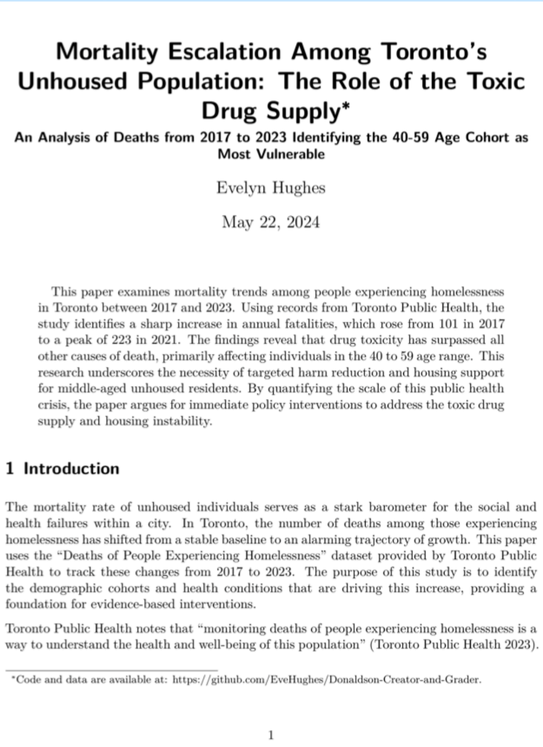
Deaths among people experiencing homelessness are a public measure of extreme social and health vulnerability. Toronto publishes a dedicated open dataset on these deaths, which creates an opportunity to describe how mortality varies over time and across demographic groupings and recorded causes. The public value of this kind of description is practical: it can help policymakers and service providers align staffing, outreach, and harm-reduction capacity with patterns in the timing and composition of deaths.

This paper addresses the following research question: **How did deaths among people experiencing homelessness in Toronto change over 2017–2023, and what patterns are visible in month-of-death seasonality, demographics (age group and gender),**

*Code and data are available at: https://github.com/EveHughes/Donaldson-Creator-and-Grader.

1

(b) GPT-5.2, Project 2 (85%)



Mortality Escalation Among Toronto's Unhoused Population: The Role of the Toxic Drug Supply*

An Analysis of Deaths from 2017 to 2023 Identifying the 40-59 Age Cohort as Most Vulnerable

Evelyn Hughes

May 22, 2024

This paper examines mortality trends among people experiencing homelessness in Toronto between 2017 and 2023. Using records from Toronto Public Health, the study identifies a sharp increase in annual fatalities, which rose from 101 in 2017 to a peak of 223 in 2021. The findings reveal that drug toxicity has surpassed all other causes of death, primarily affecting individuals in the 40 to 59 age range. This research underscores the necessity of targeted harm reduction and housing support for middle-aged unhoused residents. By quantifying the scale of this public health crisis, the paper argues for immediate policy interventions to address the toxic drug supply and housing instability.

1 Introduction

The mortality rate of unhoused individuals serves as a stark barometer for the social and health failures within a city. In Toronto, the number of deaths among those experiencing homelessness has shifted from a stable baseline to an alarming trajectory of growth. This paper uses the "Deaths of People Experiencing Homelessness" dataset provided by Toronto Public Health to track these changes from 2017 to 2023. The purpose of this study is to identify the demographic cohorts and health conditions that are driving this increase, providing a foundation for evidence-based interventions.

Toronto Public Health notes that "monitoring deaths of people experiencing homelessness is a way to understand the health and well-being of this population" (Toronto Public Health 2023).

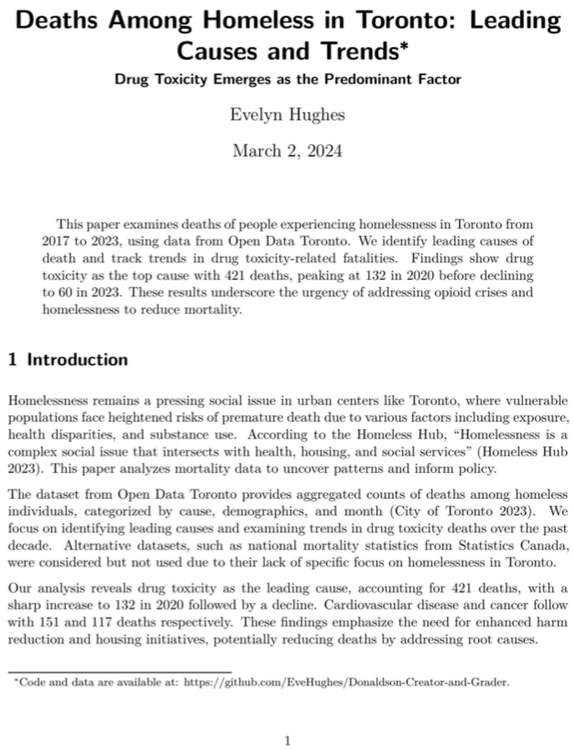
*Code and data are available at: https://github.com/EveHughes/Donaldson-Creator-and-Grader.

1

(c) Gemini 3 Flash, Project 3 (84%)

Example generated papers

The bad



Deaths Among Homeless in Toronto: Leading Causes and Trends*

Drug Toxicity Emerges as the Predominant Factor

Evelyn Hughes

March 2, 2024

This paper examines deaths of people experiencing homelessness in Toronto from 2017 to 2023, using data from Open Data Toronto. We identify leading causes of death and track trends in drug toxicity-related fatalities. Findings show drug toxicity as the top cause with 421 deaths, peaking at 132 in 2020 before declining to 60 in 2023. These results underscore the urgency of addressing opioid crises and homelessness to reduce mortality.

1 Introduction

Homelessness remains a pressing social issue in urban centers like Toronto, where vulnerable populations face heightened risks of premature death due to various factors including exposure, health disparities, and substance use. According to the Homeless Hub, "Homelessness is a complex social issue that intersects with health, housing, and social services" (Homeless Hub 2023). This paper analyzes mortality data to uncover patterns and inform policy.

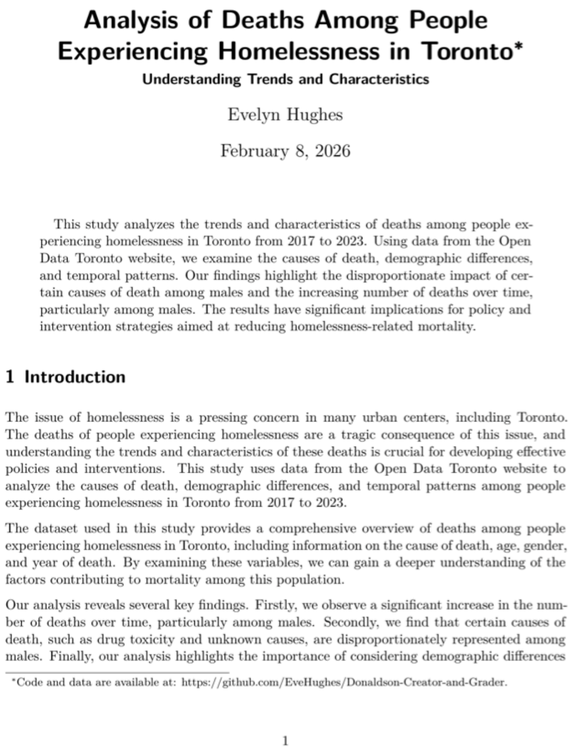
The dataset from Open Data Toronto provides aggregated counts of deaths among homeless individuals, categorized by cause, demographics, and month (City of Toronto 2023). We focus on identifying leading causes and examining trends in drug toxicity deaths over the past decade. Alternative datasets, such as national mortality statistics from Statistics Canada, were considered but not used due to their lack of specific focus on homelessness in Toronto.

Our analysis reveals drug toxicity as the leading cause, accounting for 421 deaths, with a sharp increase to 132 in 2020 followed by a decline. Cardiovascular disease and cancer follow with 151 and 117 deaths respectively. These findings emphasize the need for enhanced harm reduction and housing initiatives, potentially reducing deaths by addressing root causes.

*Code and data are available at: https://github.com/EveHughes/Donaldson-Creator-and-Grader.

1

(d) Grok-4, Project 5 (73%)



Analysis of Deaths Among People Experiencing Homelessness in Toronto*

Understanding Trends and Characteristics

Evelyn Hughes

February 8, 2026

This study analyzes the trends and characteristics of deaths among people experiencing homelessness in Toronto from 2017 to 2023. Using data from the Open Data Toronto website, we examine the causes of death, demographic differences, and temporal patterns. Our findings highlight the disproportionate impact of certain causes of death among males and the increasing number of deaths over time, particularly among males. The results have significant implications for policy and intervention strategies aimed at reducing homelessness-related mortality.

1 Introduction

The issue of homelessness is a pressing concern in many urban centers, including Toronto. The deaths of people experiencing homelessness are a tragic consequence of this issue, and understanding the trends and characteristics of these deaths is crucial for developing effective policies and interventions. This study uses data from the Open Data Toronto website to analyze the causes of death, demographic differences, and temporal patterns among people experiencing homelessness in Toronto from 2017 to 2023.

The dataset used in this study provides a comprehensive overview of deaths among people experiencing homelessness in Toronto, including information on the cause of death, age, gender, and year of death. By examining these variables, we can gain a deeper understanding of the factors contributing to mortality among this population.

Our analysis reveals several key findings. Firstly, we observe a significant increase in the number of deaths over time, particularly among males. Secondly, we find that certain causes of death, such as drug toxicity and unknown causes, are disproportionately represented among males. Finally, our analysis highlights the importance of considering demographic differences

*Code and data are available at: https://github.com/EveHughes/Donaldson-Creator-and-Grader.

1

(e) Llama 4, Project 5 (47%)

"'

title: "Analyzing the Causes and Trends of Homeless Deaths in Toronto" subtitle: "Exploring Patterns in Homeless Mortality from 2017 to 2023" author: - Evelyn Hughes thanks: "Code and data are available at: https://github.com/EveHughes/Donaldson-Creator-and-Grader." date: 28 October 2023 date-format: long

execute: echo: false

abstract: "This paper analyzes the causes and patterns of deaths among people experiencing homelessness in Toronto from 2017 to 2023. Using data from Open Data Toronto, we explore demographic trends and identify the leading causes of death. We find cardiovascular disease and drug toxicity as prominent causes, with significant demographic disparities in age and gender. This research highlights critical areas for public health intervention."

format: pdf: fig-pos: "H" number-sections: true bibliography: references.bib —

Introduction

Homelessness is a significant social issue, affecting the health and well-being of thousands. Understanding mortality trends in this population is crucial for public health planning. In this paper, we analyze data from Open Data Toronto focusing on the deaths of homeless individuals. Our goal is to identify key patterns and causes of death that may inform public health interventions. Through a detailed examination of age, gender, and cause of death, we provide insights into potential preventive measures.

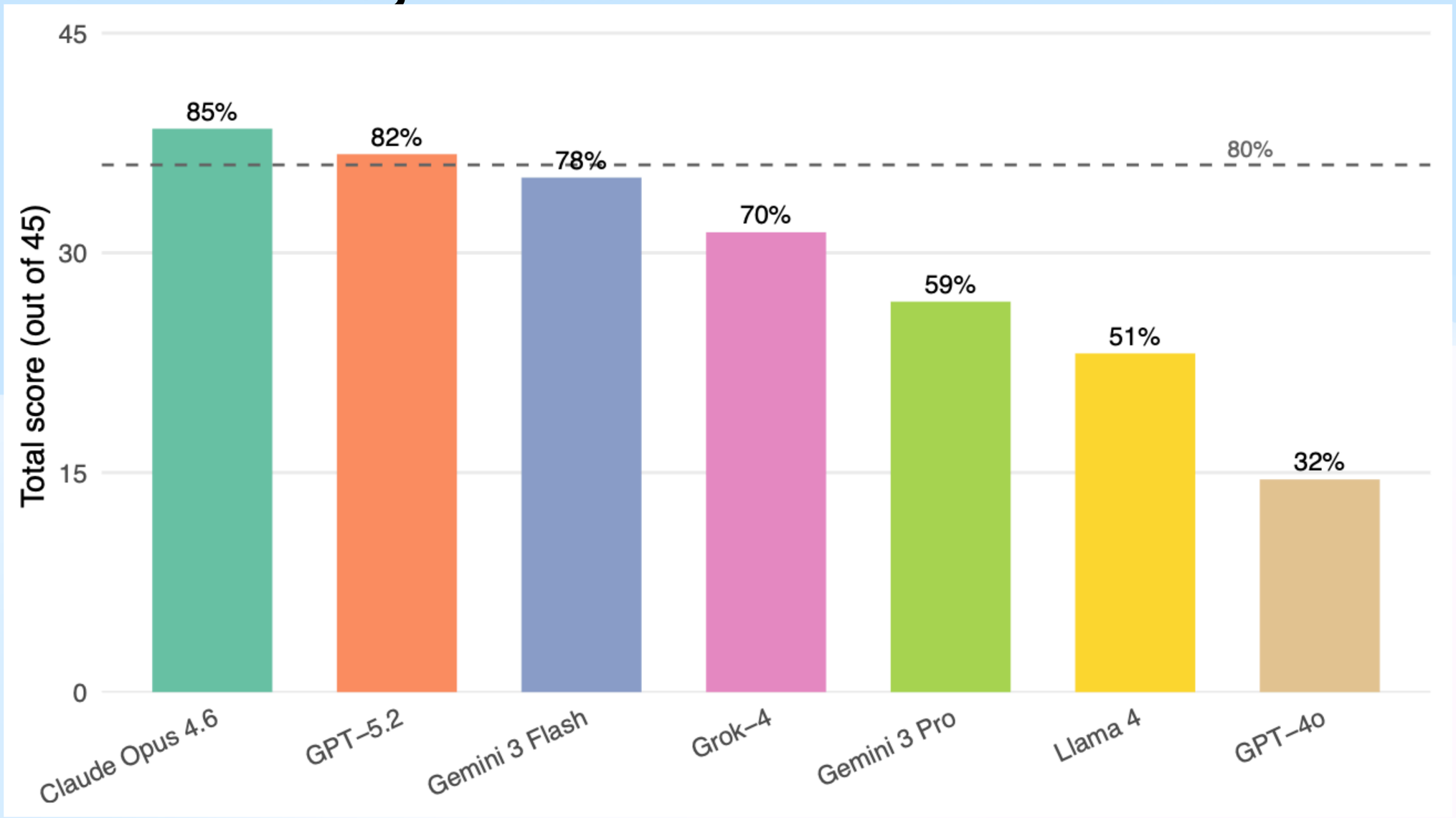
Data

The dataset, "Deaths of People Experiencing Homelessness," was sourced from Open Data Toronto [@toronto_homeless_deaths]. It contains records from 2017 to 2023, detailing each death with attributes such as cause, age group, and gender. Similar datasets were considered but were not as comprehensive in terms of demographic details. The primary variables of interest include `Year of death`, `Cause_of_death`, `Age_group`, and `Gender`.

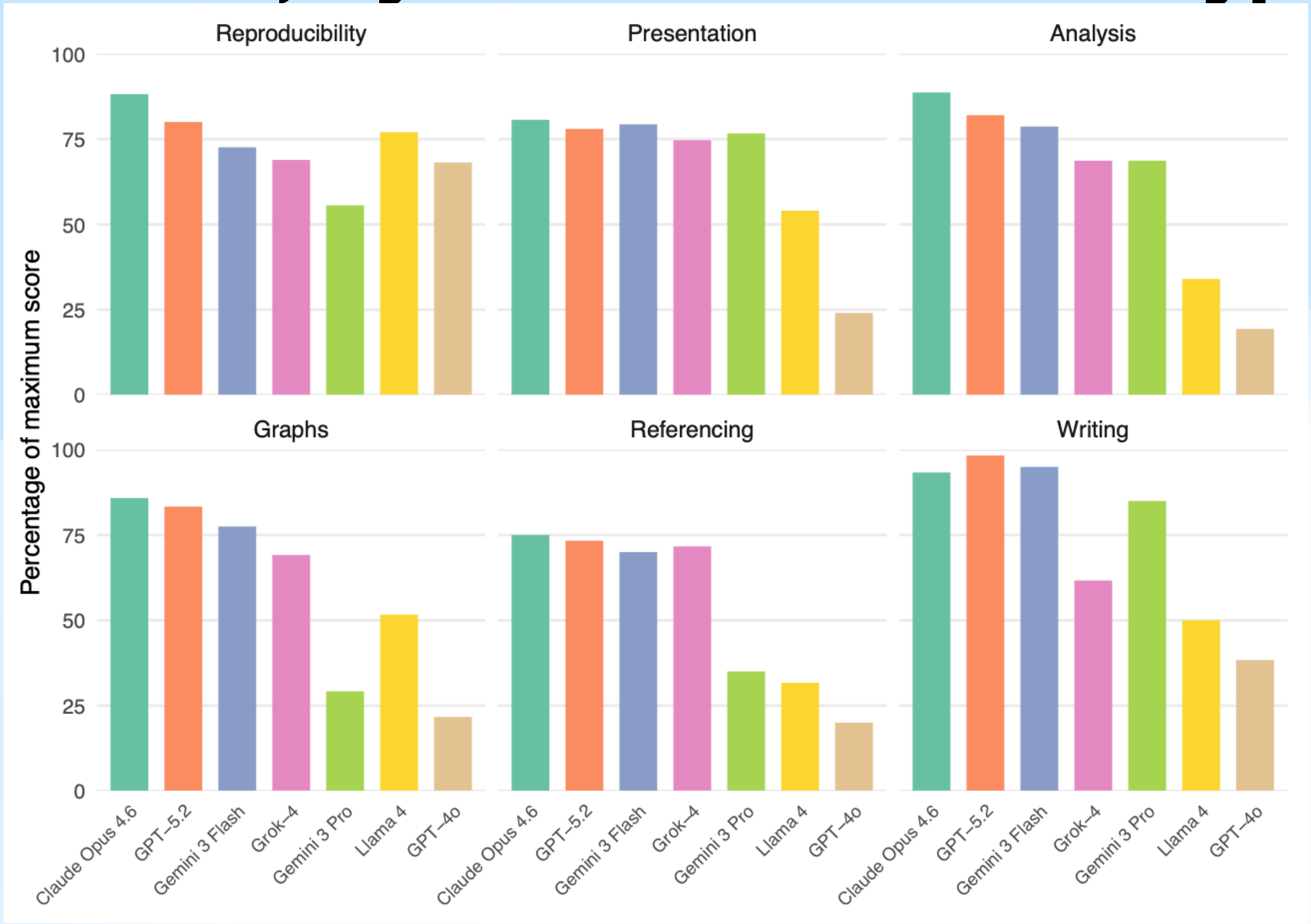
1

(f) GPT-4o, Project 2 (13%)

Performance, overall

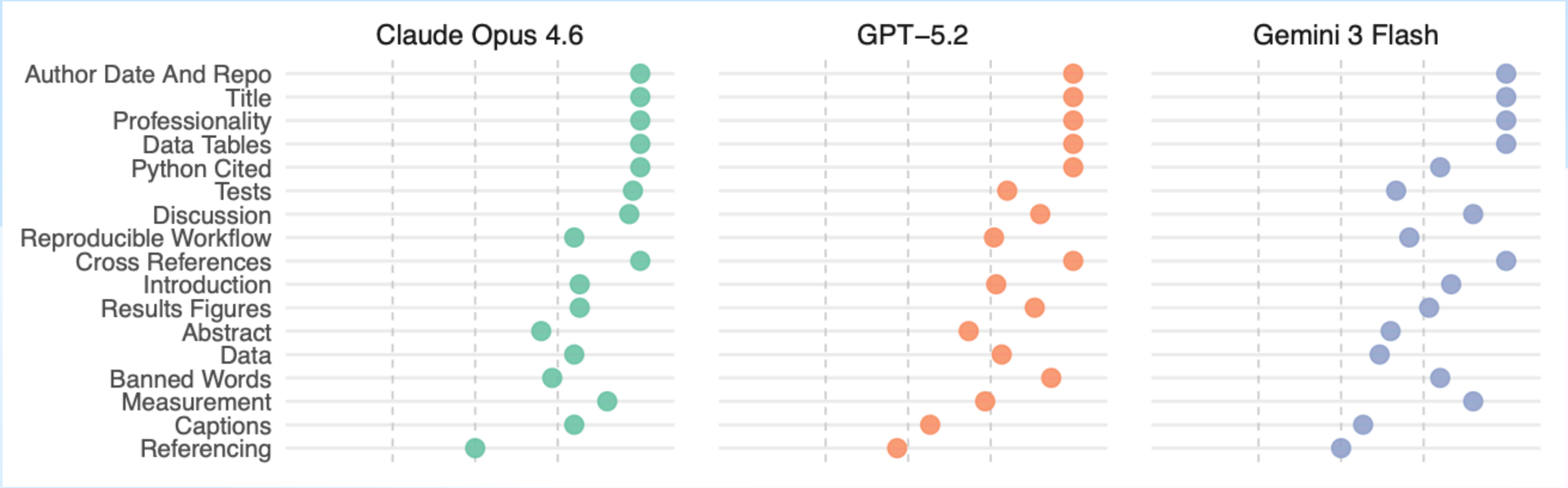


Performance, by rubric evaluation type



Performance, by rubric category

Three best performing models



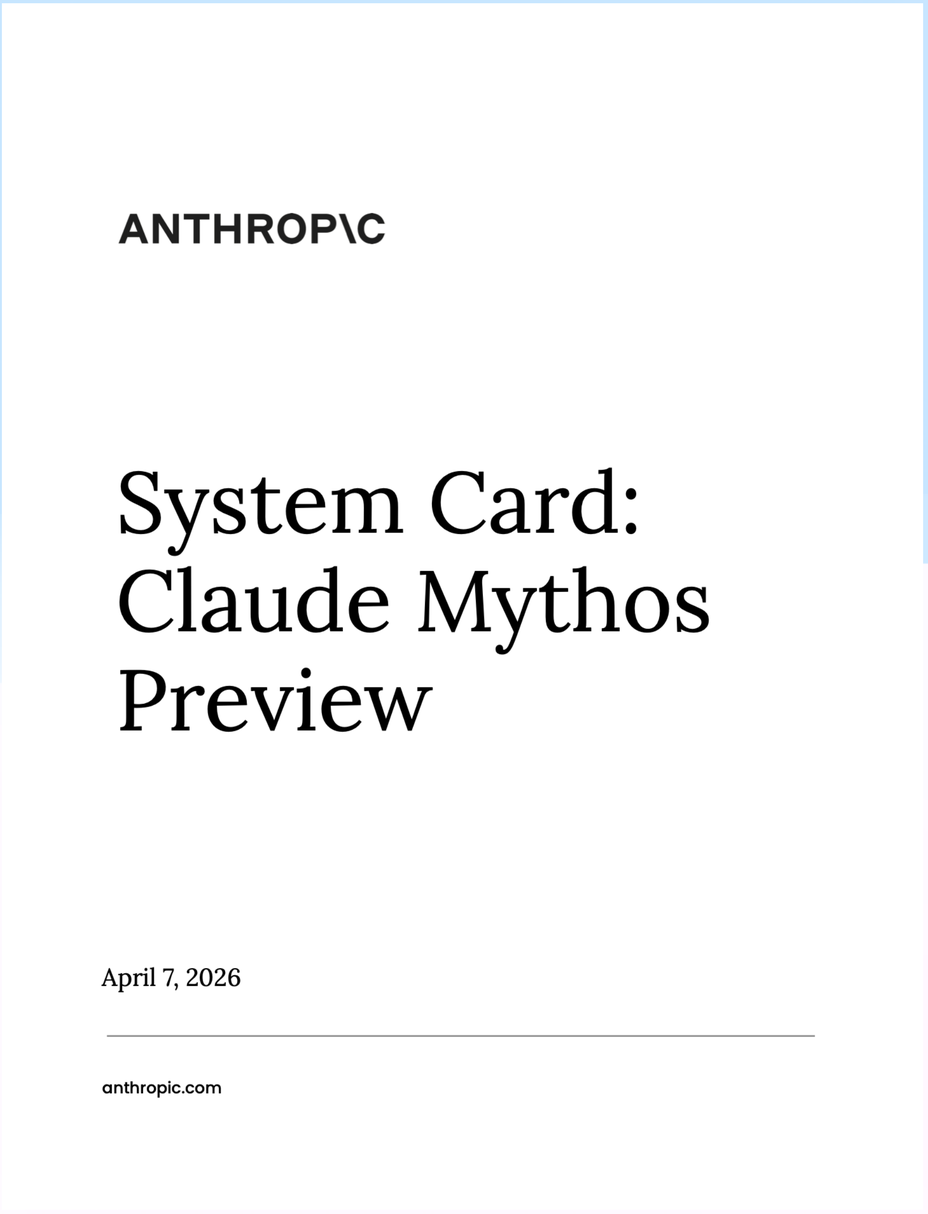
Key takeaways

- Even the best current frontier models still require human oversight.
- The capacity to generate high-quality end-to-end data science projects differs across frontier models.
- But comparing two models from OpenAI, one on the frontier and the other a year old, it is possible that continued development could change this in the next few years.
- We need considerably more domain-specific benchmarks.
- Getting zero on cross-references suggests a failure to understand Quarto syntax even when there were explicit instructions and examples.
- The best models scored similar to an A-/B+ undergraduate mark.

But don’t take my word for it!

“...overconfidence and poor strategic judgment in distinguishing high-value ideas from implausible ones”

(Mythos model card, p. 26)



What has changed?

In the age of AI

1. The importance of teaching skepticism.
    - Evals rule everything around me.
2. The importance of teaching editing and understanding code
    - “Learning Opportunities: A Claude Code and Codex Skill for Deliberate Skill Development” by Cat Hicks: https://github.com/DrCatHicks/learning-opportunities.
3. The importance of teaching students to push back.
    - AI helps most where the user knows least (Mythos model card, p. 27)!

Concluding remarks

“May you live in interesting times”

It was her voice that made
The sky acutest at its vanishing.
She measured to the hour its solitude.
She was the single artificer of the world
In which she sang. And when she sang, the sea,
Whatever self it had, became the self
That was her song, for she was the maker.

Extract “The Idea of Order at Key West”, Wallace Stevens, 1934.