Using LLMs for Science

Jared Donohue

https://jareddonohue.com/

My Story

- B.S. in Computer Science (2016)
  - Economics → Computational Data Science → Computer Science
  - Thought about academic route late; got interviews at big tech -- Google, Facebook, Amazon → Accepted Amazon in Boston to work on Alexa
- Worked at Amazon for 8 years (2017-2024)
  - 4yrs SDE → 4yrs PM (customer-facing products, research, experimentation)
- M.S. in Data Science (2025)
  - Interested in experimentation, causal inference, behavior from PM work
  - wanted to try research then consider PhD after
- Now: R.A. at Columbia Business School (2026) experimentation and causal inference projects
  - using LLMs for “science” every day
  - deciding industry vs PhD for 2027

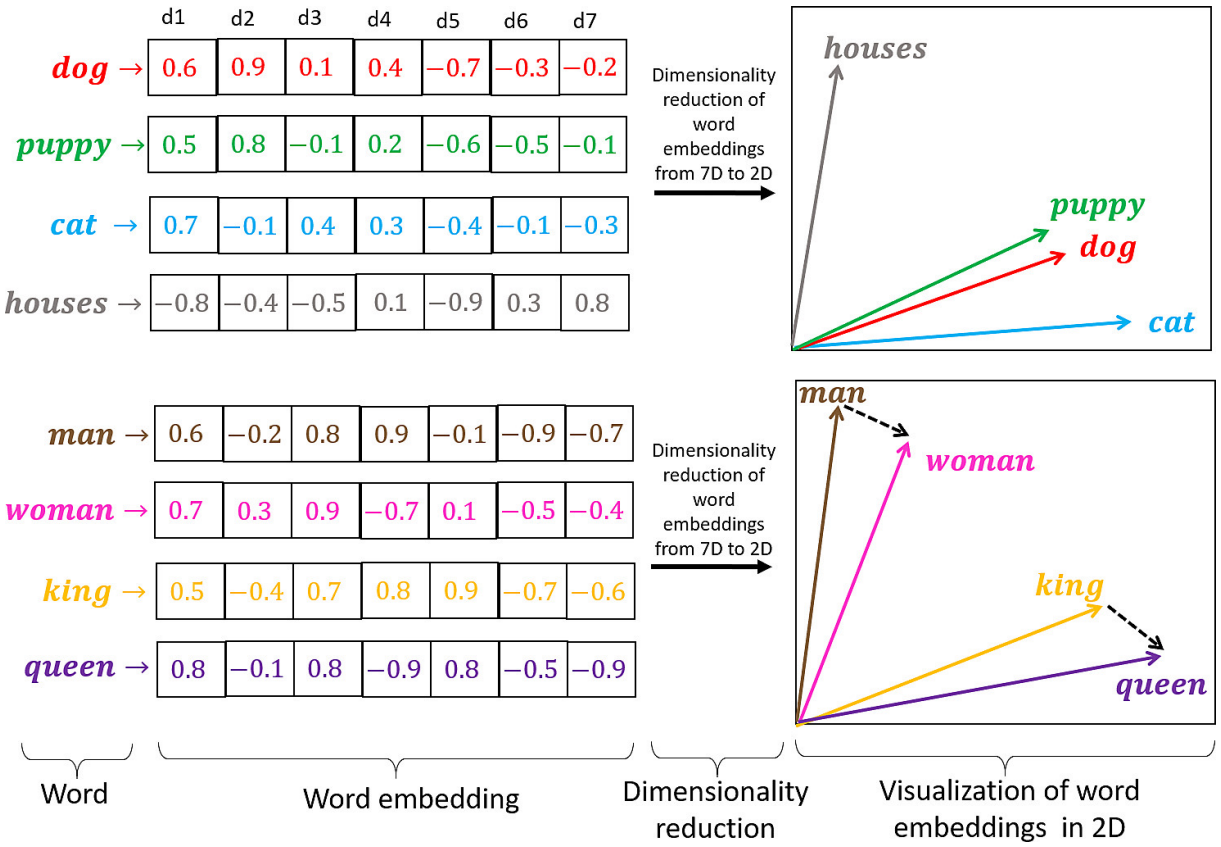
Using LLMs for Science

Jared Donohue

https://jareddonohue.com/

What are LLMs?

From an LLM, with framing: *Large Language Models (LLMs) are* ***neural network stat models trained*** *on* ***massive text corpora*** *that use* ***embeddings***, ***transformers***, *and* ***attention*** *mechanisms to predict and* ***generate statistically likely text*** *given an input prompt.*



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

Fun Exercise: Building a Sentence Like an LLM

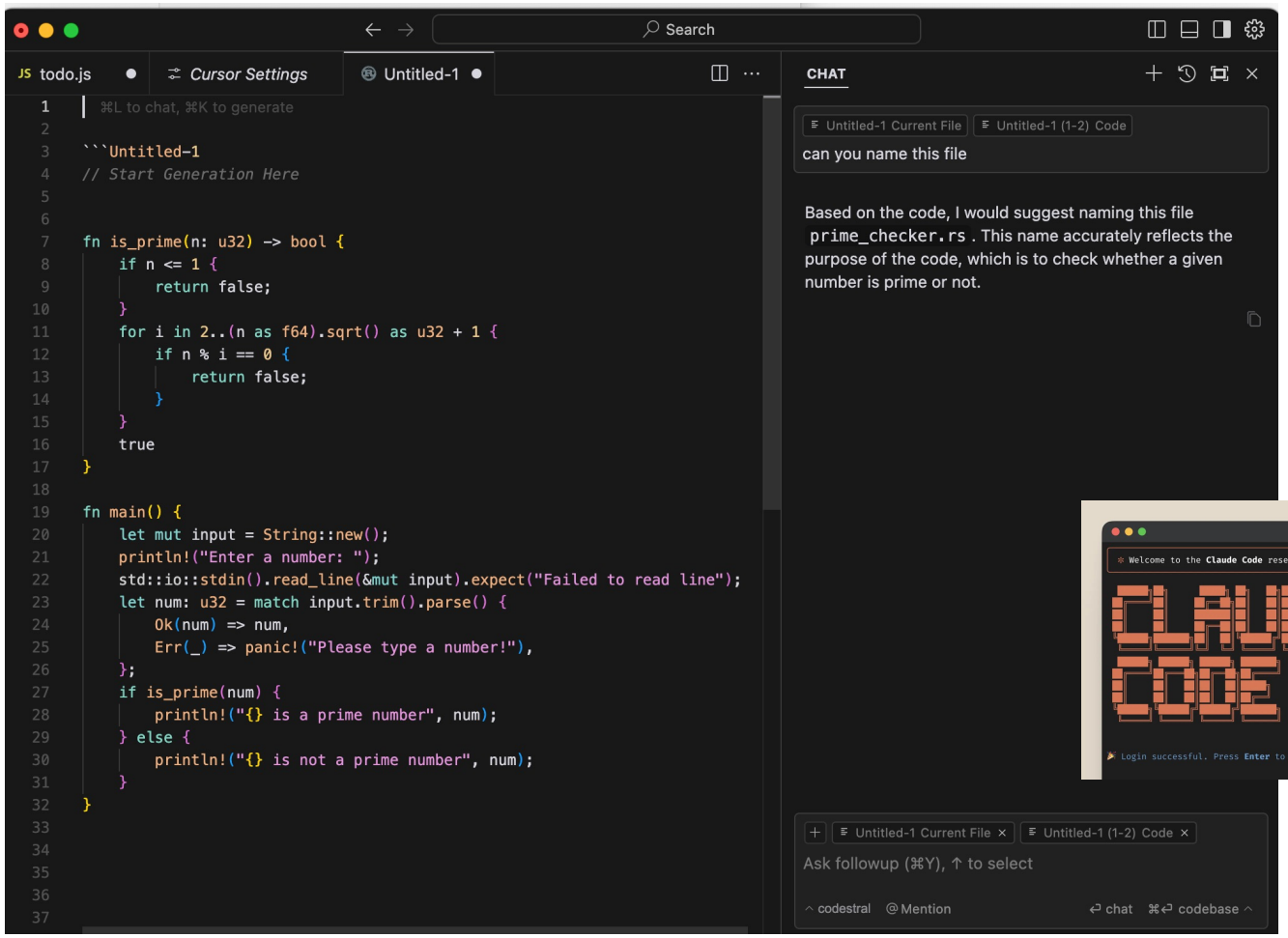
- *Good analogy for how an LLM works → generating text one word (token) at a time by predicting the most likely next word from context.*

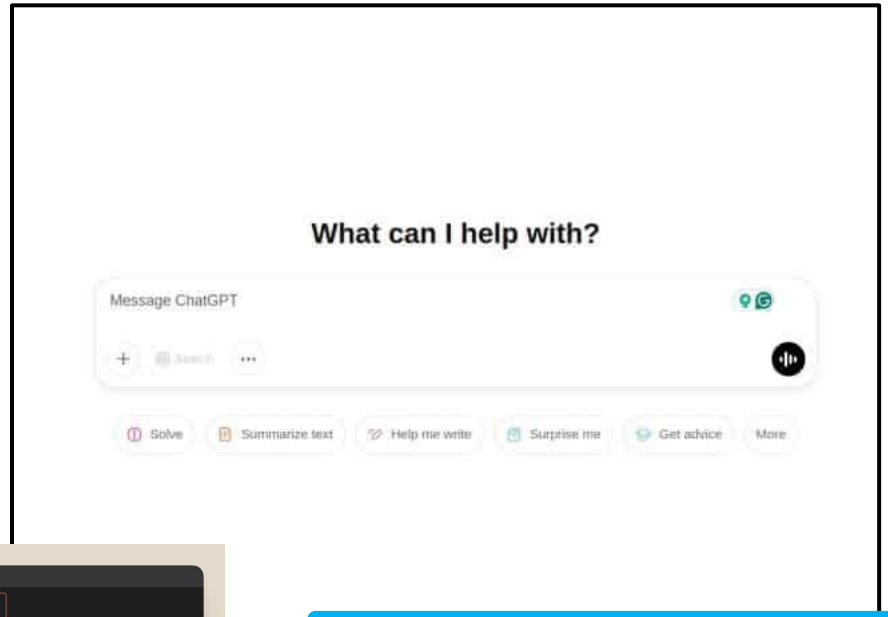
LLMs as a Tool

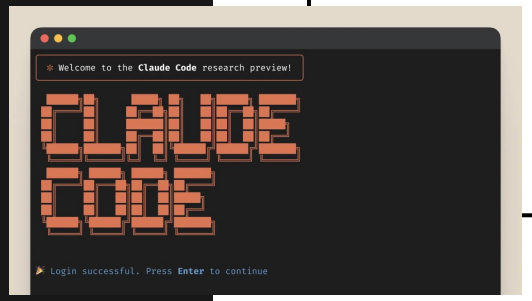
- LLMs for Science = **LLMs as a tool, very good at generating sensible text and reasoning (maybe)**, given sensible input

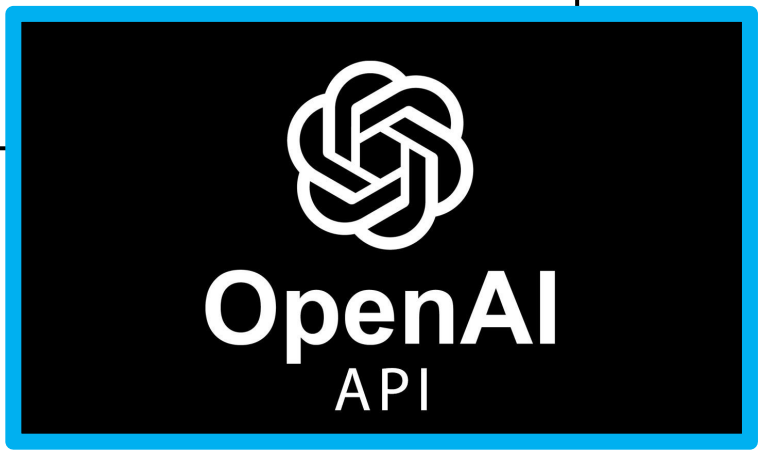
- **Great for coding**.
  - Quickly verifiable
  - Lots of high quality training data (stack overflow, GitHub)
  - How do you use LLMs for coding? For learning? Which LLMs?

LLM Tools: IDE, Chat, API, and more





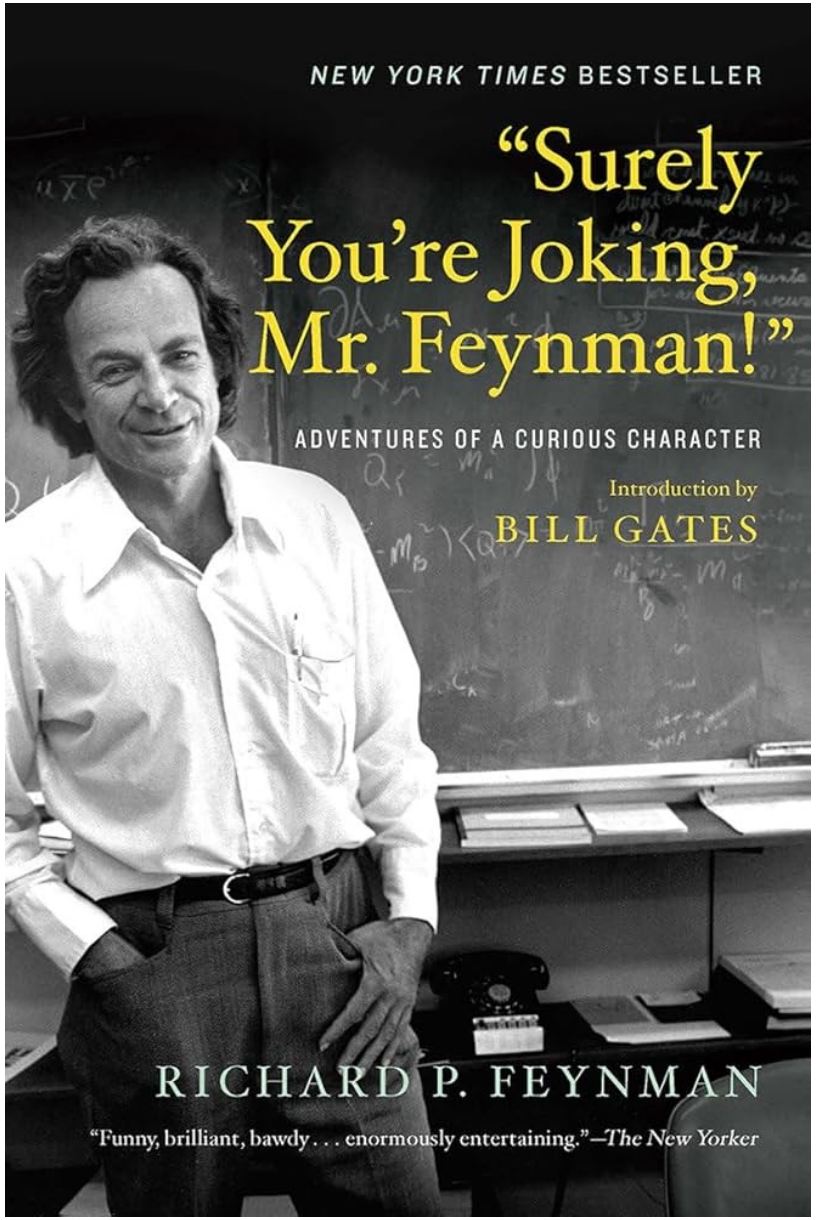




What is Science?

- Philosophy of Science
  - Discovering true **“laws” of nature**
  - **Problem of induction** (all empirical science). Can never observe everything. Assumes **“uniformity of nature”**

- According to Feynman…
  - *“…*to **doubt that what is being passed from the past is in fact true, and to try to find out** ab initio (“from the start”) again from experience…*the result of the discovery that it is* ***worthwhile rechecking by new direct experience****, and not necessarily trusting the human race’s experience from the past…”*
  - https://feynman.com/science/what-is-science/

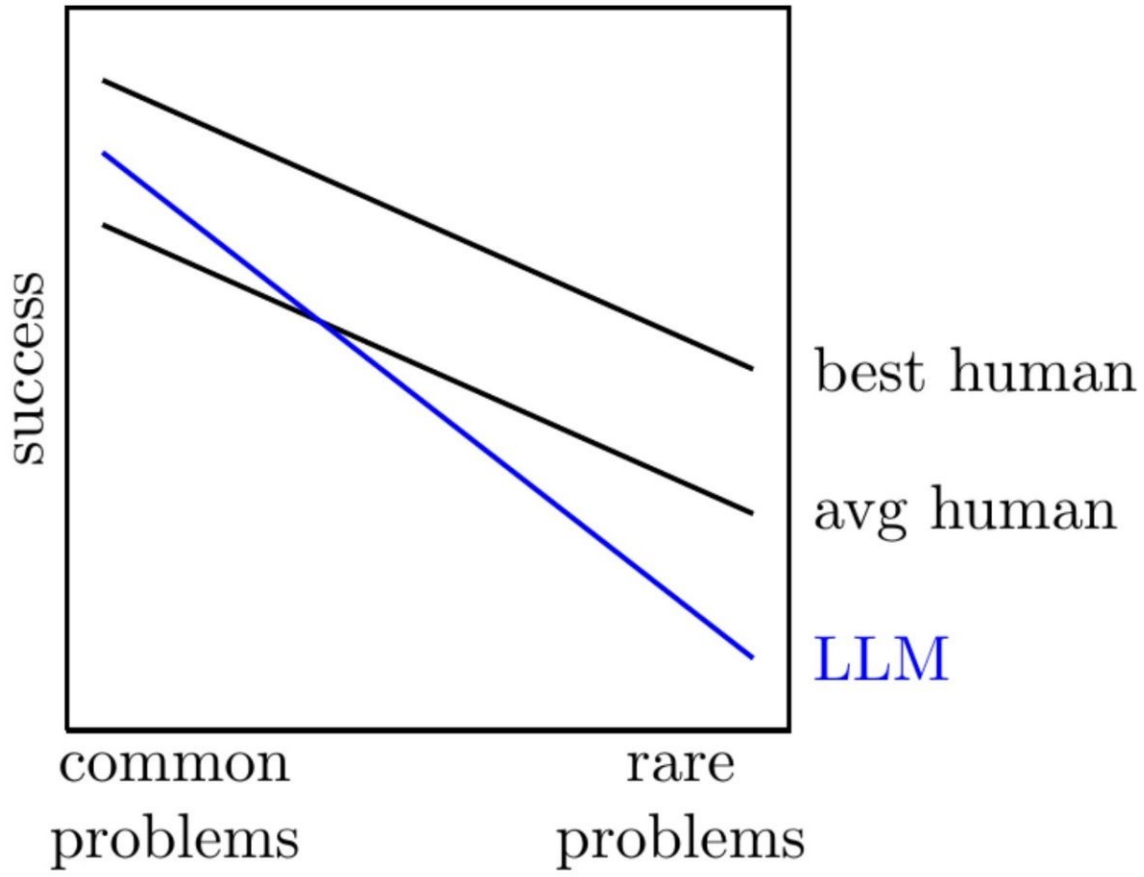


Typical Scientific Paper (as I see it)

- **Modern science involves producing papers which document a contribution to a body of knowledge in a field** (e.g. Physics, Earth and Environmental Science)

- Typical Scientific Paper
  - **Theory** – deep domain knowledge and synthesis of ideas
  - **Data** – requires data collection or experiment
  - **Empirical Analysis** – writing code (e.g. Python) to analyze data
  - **Narrative** to explain, usually writing LaTeX

Aside: AI Scientists and AI vs Human by Problem Type



AI Scientists

- https://sakana.ai/ai-scientist/
- https://agents4science.stanford.edu/

Back to: **Using LLMs for Science**

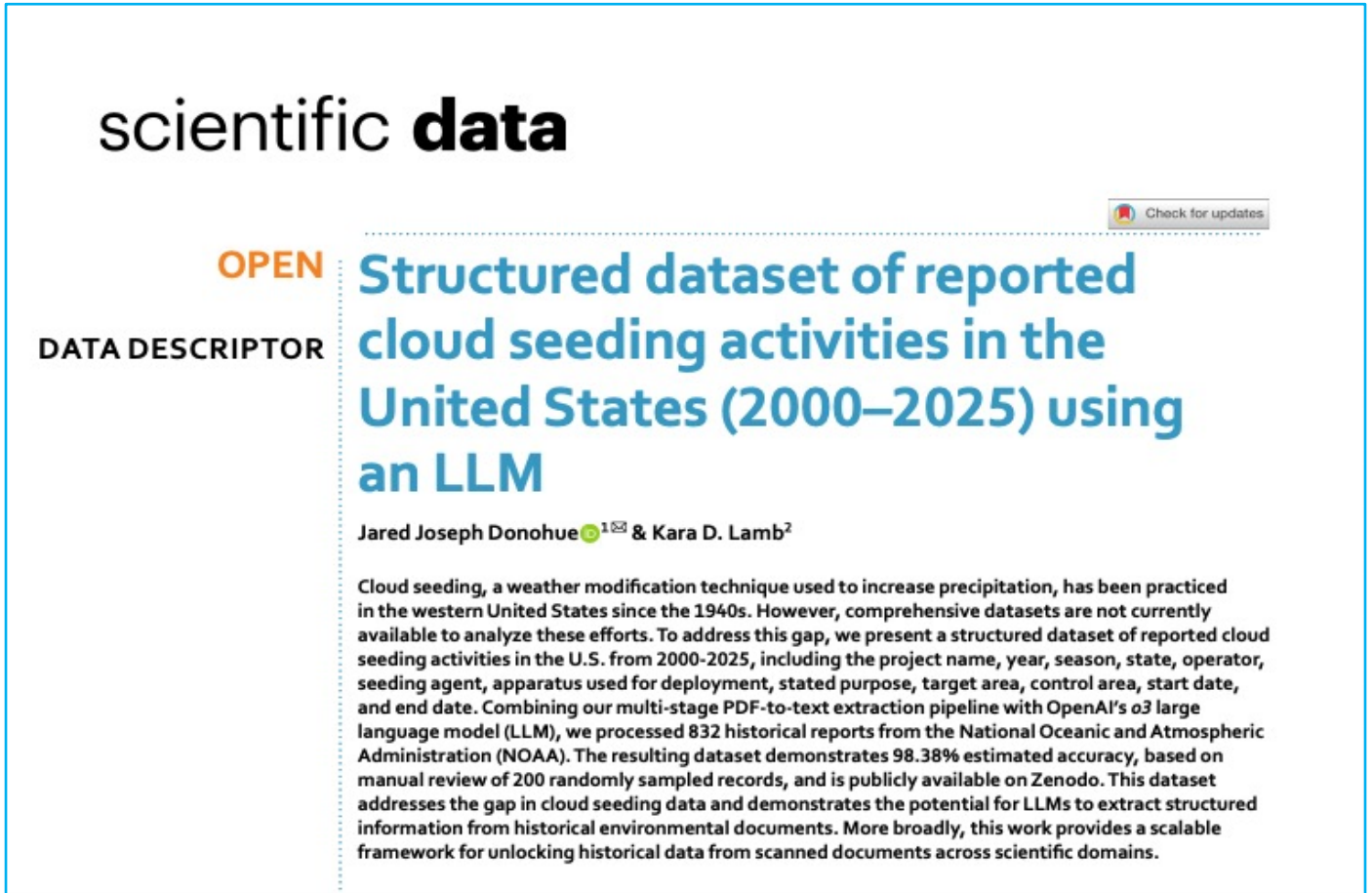
- We want to **leverage LLM tools for producing a scientific paper with theory, data, empirical analysis, and narrative,** which adds to a body of knowledge in a field (e.g. atmospheric science)

- Scientific Paper
  - **Theory** – deep domain knowledge and synthesis of ideas; **LLMs BAD**
  - **Data** – requires data collection or experiment; **LLMs BAD**
  - **Empirical Analysis** – writing code (e.g. Python) to analyze data; **LLMs GOOD**
  - **Narrative** to explain, usually writing LaTeX; **LLMs GOOD**

- My experience using LLMs for Science …

Structured dataset of reported cloud seeding activities in the United States (2000–2025) using an LLM

Jared Donohue (me) and Kara D. Lamb (shoutout / thank you)
https://www.nature.com/articles/s41597-025-06273-1



scientific data

OPEN

DATA DESCRIPTOR

Check for updates

Structured dataset of reported cloud seeding activities in the United States (2000–2025) using an LLM

Jared Joseph Donohue[1]✉ & Kara D. Lamb[2]

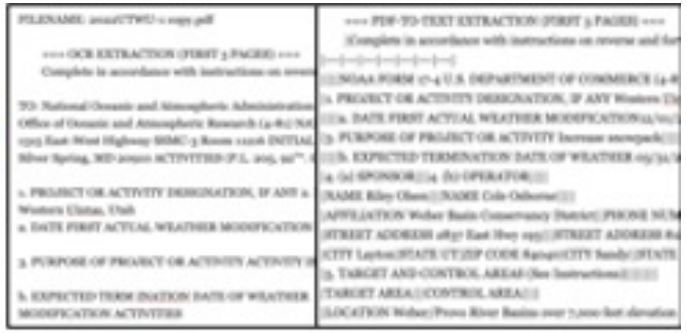
Cloud seeding, a weather modification technique used to increase precipitation, has been practiced in the western United States since the 1940s. However, comprehensive datasets are not currently available to analyze these efforts. To address this gap, we present a structured dataset of reported cloud seeding activities in the U.S. from 2000-2025, including the project name, year, season, state, operator, seeding agent, apparatus used for deployment, stated purpose, target area, control area, start date, and end date. Combining our multi-stage PDF-to-text extraction pipeline with OpenAI's *o3* large language model (LLM), we processed 832 historical reports from the National Oceanic and Atmospheric Administration (NOAA). The resulting dataset demonstrates 98.38% estimated accuracy, based on manual review of 200 randomly sampled records, and is publicly available on Zenodo. This dataset addresses the gap in cloud seeding data and demonstrates the potential for LLMs to extract structured information from historical environmental documents. More broadly, this work provides a scalable framework for unlocking historical data from scanned documents across scientific domains.

How LLMs were used: text extraction/reasoning

Form 17-4

Form 17-4A

convert PDF to text



prompt engineering

parse LLM response and save to CSV

	A	B	C	D	E	F	G	H	I
1	filename	start_date	end_date	season	year	state	agent	apparatus	purpose
2	2018CONBWC_vailbeaver	11/1/18	3/31/19	winter	2018	colorad	silver iodide, sodiun	ground	winter snow pack and pr
3	Panhandle Ground water	4/15/03	9/30/03	spring	2003	texas	silver iodide	airborne	precipitation enhancem
4	Southern - Central Utah_	11/1/09	4/15/10	winter	2009	utah	silver iodide	ground	augment snowpack
5	North Dakota Cloud Mod	6/1/11	8/31/11	summer	2011	north d	silver iodide	ground	rainfall increase and hail
6	2022UTS-1.pdf	11/1/22	12/24/22	winter	2022	utah	silver iodide	ground	augment early season sn
7	2024CASAWPA-1.pdf	11/15/23	12/31/23	fall	2024	californ	silver iodide	airborne	increase precipitation
8	DWD - Upper Ark. Prograr	11/9/10	4/10/11	winter	2010	colorad	silver iodide, sodiun	ground	cloud seeding for precipi
9	Kaweah River Project_00	1/1/00	12/31/00	winter	2000	californ	silver iodide	ground, airbo	precipitation augmentat
10	2023WYWR-F.pdf	11/15/20	4/15/21	winter	2021	wyomin	urea, carbon dioxid	airborne	weather modification act

How LLMs were used: text extraction/reasoning

- **Input**: PDFs sent to NOAA on Cloud Seeding activities (req'd by law)
  - Pymu4llm – text-based PDFs
  - Tesseract (optical character recognition - OCR) – scanned PDFs
  - *Unstract* (https://unstract.com/llmwhisperer/) – both and formatting but $
- **Prompt**
  - Reasoning models best (e.g. o3)
  - Using examples/reasoning in the prompt
- **Output**
  - https://zenodo.org/records/16754931 (CSV)

LLM Prompt Excerpt (example + reasoning)

Internal Reasoning Style (Do Not Include in Output)

Use structured internal logic to disambiguate each field. For example:

YEAR:

- Filename: "2018UTNORT-1.pdf" → clearly indicates 2018
- Text also mentions winter 2017–2018 → confirm choice of later year
- Final value: **2018**

STATE:

- Filename: "UT" and text: "northern utah"
- Final value: **utah**

Historical Analysis of Cloud Seeding in the U.S. using LLMs

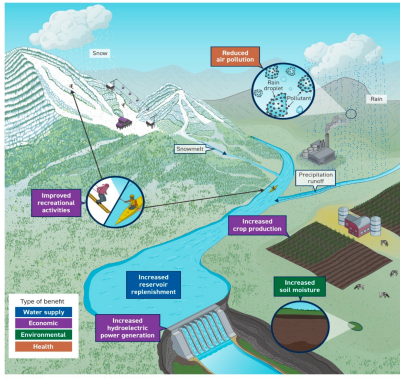
Data Science Institute
COLUMBIA UNIVERSITY

Jared Joseph Donohue
Kara D. Lamb

DSI Scholars Project
Kara D. Lamb, Earth and
Environmental Engineering

Topic

Cloud seeding is a weather modification technique used to enhance precipitation using chemical agents. Despite its use in the Western U.S. since the 1940s, historical data for cloud seeding activity in the U.S. is hard to find. We address this gap by extracting structured records from historic NOAA reports (2000–2025) required by law since 1972.



Source: GAO (analysis and illustration). | GAO-25-107328

Figure 1. Cloud Seeding Ecosystem

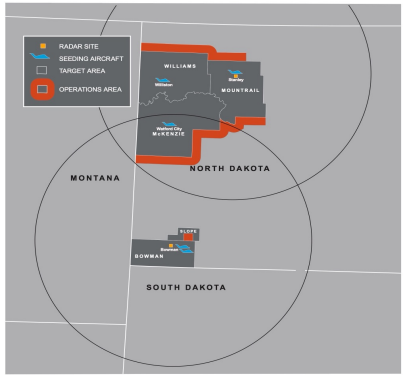


Figure 2. North Dakota Cloud Seeding Program

Methods

We used multi-stage PDF-to-text conversion and chain-of-thought LLM prompting to process 836 historical NOAA reports and store 9 metadata fields into CSV.

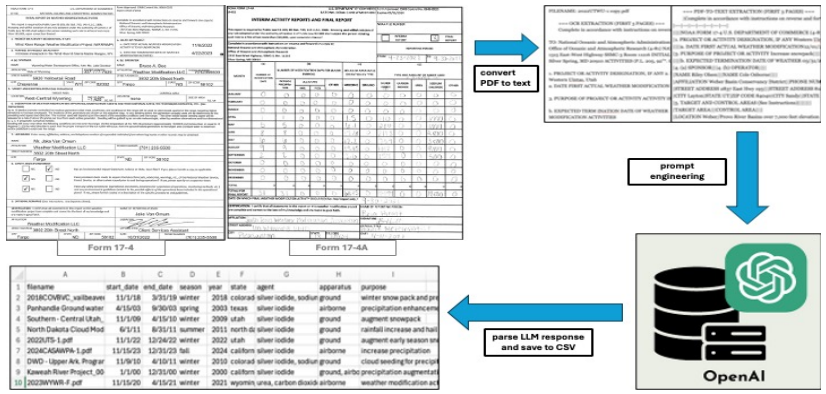


Figure 3. System Diagram showing NOAA Forms, PDF-to-Text Conversion, OpenAI API

Results

98.38%

Our dataset achieved 94.72% average accuracy across the extracted fields. Analyzing the dataset, we find that most activity is concentrated in Utah, Colorado, and eastern California, typically for snowpack augmentation using ground-based silver iodide.

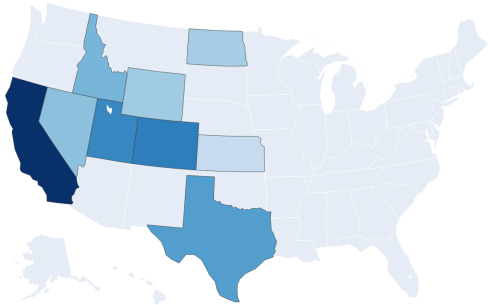


Figure 4. Cloud Seeding Activity in the U.S. by Activity Count (2000-2025)

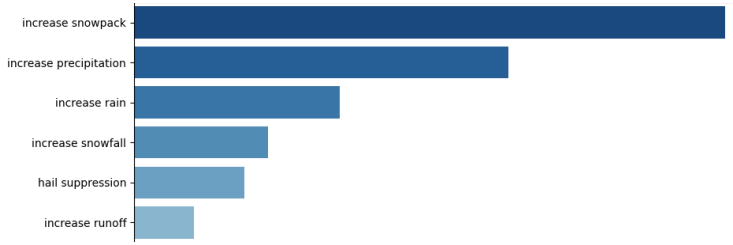


Figure 5. Count of U.S. Cloud Seeding Activities (2000-2025)

Conclusion

We addressed the data gap in U.S. weather modification records and demonstrated that LLMs, paired with chain-of-thought prompts and careful text preprocessing, can reliably extract structured environmental data from inconsistent, scanned historical reports, offering a scalable model for other scientific domains.

Acknowledgments

Kara D. Lamb for the exceptional research mentorship, and collaboration on the project itself.

References

United States Government Accountability Office (Figure 1)
North Dakota Weather Modification Program (Figure 2)
NOAA Weather Modification Project Reports (Figure 3)

Select site

Central Colorado Program (…

State: colorado

Program: Central Colorado Program

Location: 39.60°N, 253.48°W

Seeded months: 96 / 276

Unseeded months: 180 / 276

Cloud Seeding Difference-in-Differences Explorer

Target vs control precipitation. We compare the target−control gap during seeded months to the same gap during unseeded months. The aggregate ATT is the equal-weighted mean of per-site DiDs, so every site counts the same regardless of how many months it seeded. The DiD is the difference in the target-control gap between seeded and unseeded months.

Aggregate causal effect (equal-weighted across sites)

ATT (AVERAGE TREATMENT EFFECT ON THE TREATED)

+0.046 mm

Equal-weighted mean of per-site DiDs (n = 129 sites)

p-value
0.2745

95% CI
[-0.037, +0.129] mm

Std. error
0.042

Per-site causal effect estimate (within-site DiD)

WITHIN-SITE DID

+1.172 mm ***

Seeded months: 96 | Unseeded months: 180

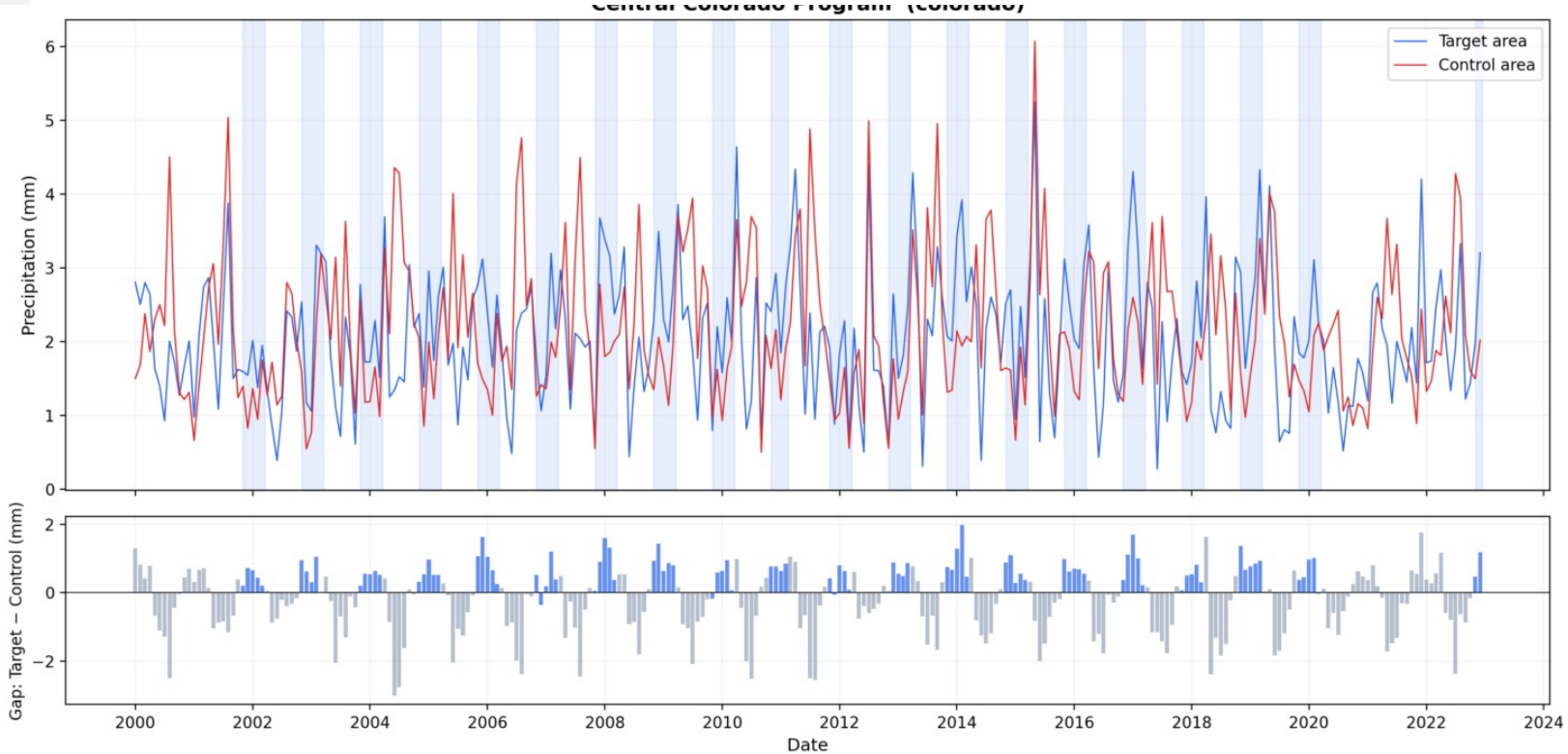
p-value
0.0000

95% CI
[+1.010, +1.333] mm

Std. error
0.082

Mean gap, unseeded months
-0.51 mm

Mean gap, seeded months
+0.66 mm



This Same Paper Without LLMs?

- OCR + REGEX rules + classic NLP + custom business logic, rather than one broadly capable text system
- Nothing as generalizable and does your specific task right away
- LLMs: far more flexible, faster to deploy, and usable on messy text immediately
- Much faster LaTeX manuscript (fix errors, generate draft paragraphs)

What I have learned using LLMs for Science (final remarks)

- **Know what you want first (coding, writing)**
  - LLMs will jump to building complex solutions, writing way too much
  - Try setting *claude.md* etc. preferences. May still need to specify in prompt

- **Know Available Tools and Technologies (coding)**
  - Great way to guide LLM to what you want / know – *Streamlit*
  - LLM not good at finding these on its own

- **Code and Verify in Incremental Steps**
  - Maintainability, Explainability, Reproducibility, Verifiability matter.
  - Know how to verify and explain to a collaborator, peer reviewer
  - Verify after each step

What I have learned using LLMs for Science (final remarks)

- **Spend the time on Input Data Quality**
  - Source documents (manual inspection+Pymu+OCR+Unstract)
  - Prompt engineering (filename had the dates example)

- **Customize Your Writing** (take pride in this being your work)
  - LLM great for producing text, even LaTeX, but cheap prose -- read papers and you can tell which are written by AI // em-dashes, “more broadly”, “not only this but that”… at least bring your writing to the prompt

- **Use the tools; think “with”; let AI improve your scientific work**

Thank you!

- Feel free to reach out, connect on LinkedIn, or ask any follow-up questions you might have.
- Email: jjd2203@columbia.edu

Q&A