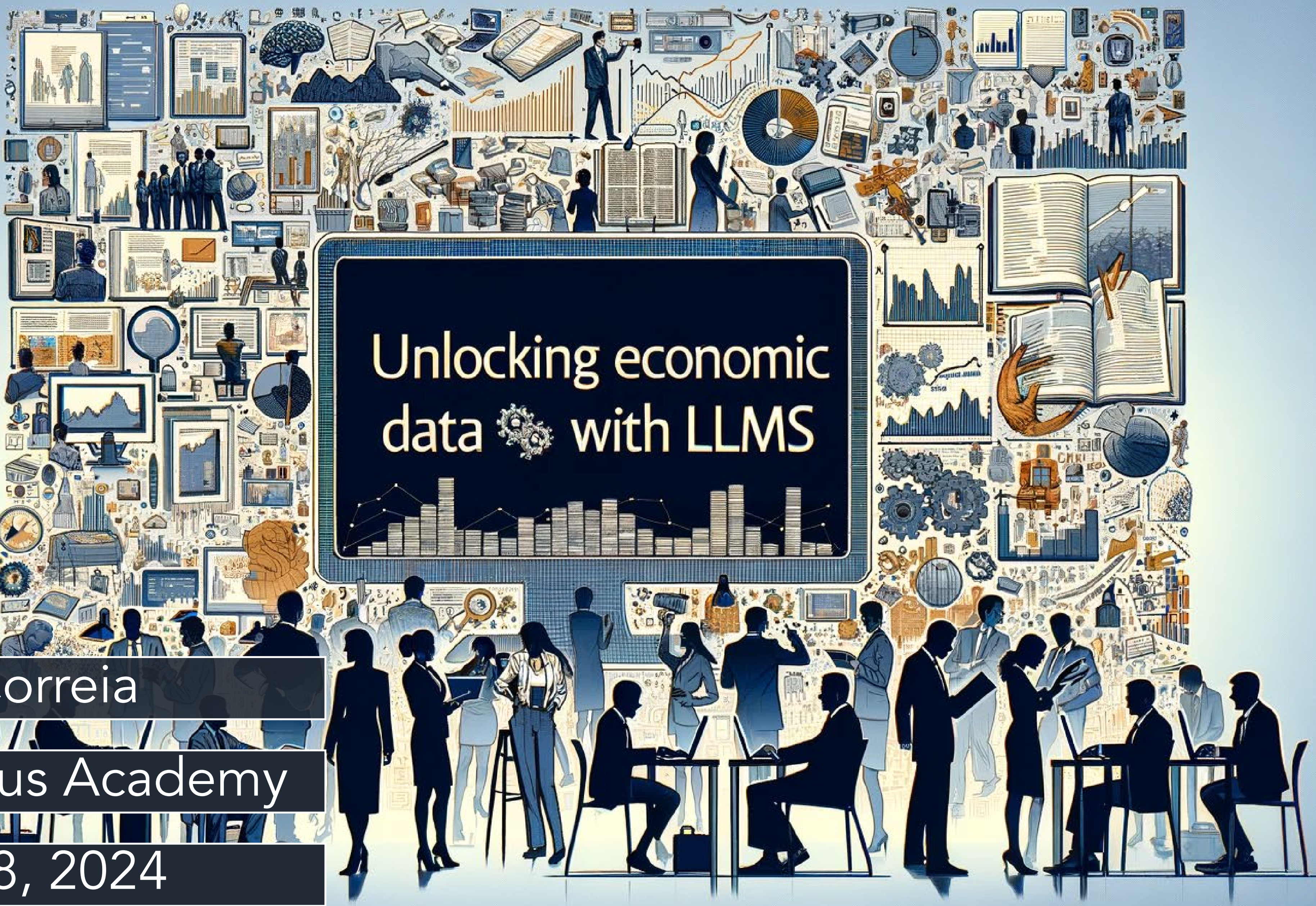
Disclaimer

This research presentation is solely the responsibility of the author and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve or any other person associated with the Federal Reserve System.

References to specific tools, programming languages, or datasets within this presentation are for illustrative purposes only and should not be taken as an endorsement or recommendation.

Any code snippets shown throughout the talk are for informational purposes and are presented without warranty. Users should review and test the code thoroughly before any reliance or implementation.

The images included in this presentation were generated using GenAI tools such as GPT and Midjourney, and are intended solely for illustrative purposes.



Sergio Correia

@ Markus Academy

March 28, 2024

This talk

- Not general advice on LLMs:
  - See Anton Korinek's 2023 JEL or Kevin Bryan @ Markus Academy for advice not specific to economics (brainstorming, proofreading, coding, etc.)

- Only about research, not policy

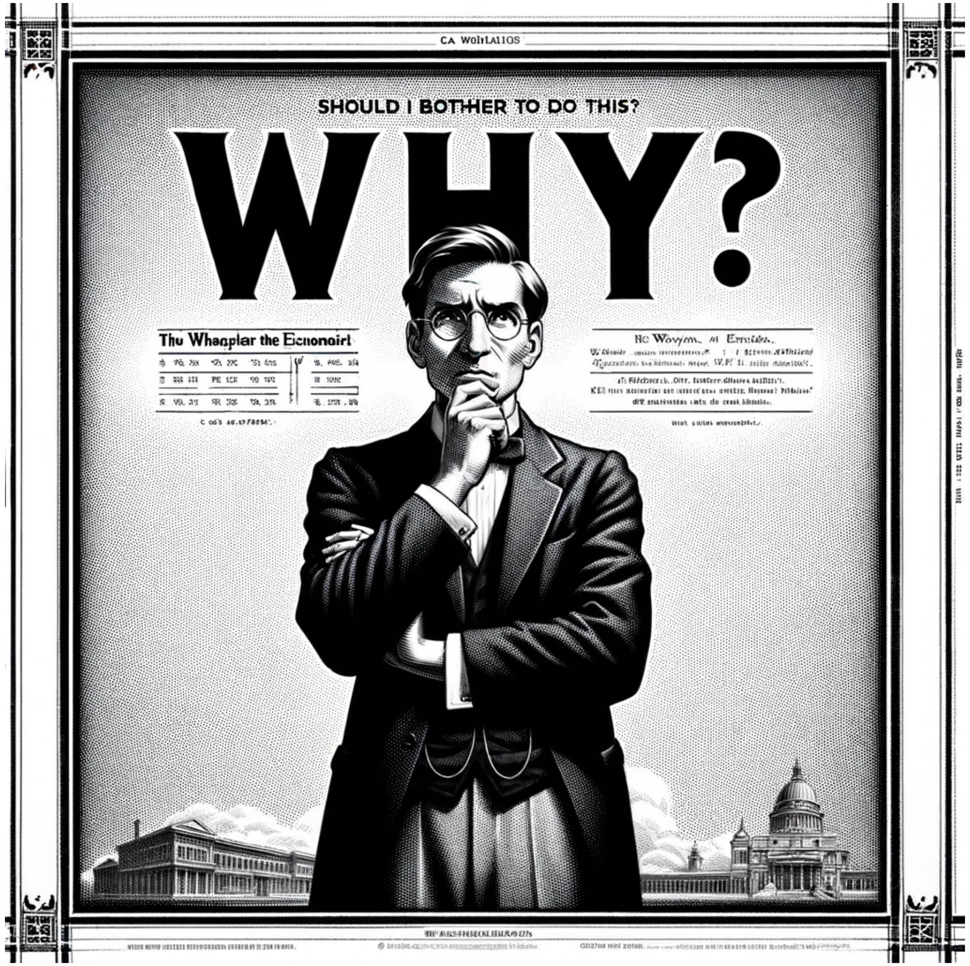
- Focus on data

This talk

1. [Why]
   LLMs can be used to generate novel data useful for research.

2. [How]
   Barriers to entry appear high, but it's actually quite simple!

Act 1:



Why?

A good empirical econ/finance paper needs:

1. An important question (= economics)
2. A clean identification strategy
3. A dataset where you can test #1 with #2

Typical PhD student mistake:

- Writing an empirical paper with just Compustat or public Call Reports
- Problem: thousands of smart people have gone through that data before; what's your edge?
- Testing a new theory? novel identification strategy? If not, uphill battle

Alternative: get a novel dataset!

- Confidential or private datasets
- Historical datasets
- And nontraditional datasets

- Doesn't solve requirements #1 and #2, but allows us to ask better questions; use cleaner identification strategies

Nontraditional datasets

ELSEVIER

Journal of Monetary Economics

Available online 15 September 2023

In Press, Corrected Proof ⓘ What's this?

More than words: Fed Chairs' communication during congressional testimonies ☆

Michelle Alexopoulos [a], Xinfen Han [b], Oleksiy Kryvtsov [c], Xu Zhang [b]

Show more

+ Add to Mendeley Share Cite

A.6 Snapshots of the Fed Chair and Congress members' face-emotions

Face: Ben Bernanke
Face emotion score: -0.222

Face: Michael Castle
Face emotion score: -0.293

Table A.4. Facial Emotions - July 22, 2010 Testimony

Nontraditional datasets

JOURNAL ARTICLE

What We Teach About Race and Gender: Representation in Images and Text of Children's Books*

Anjali Adukia, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, Teodora Szasz

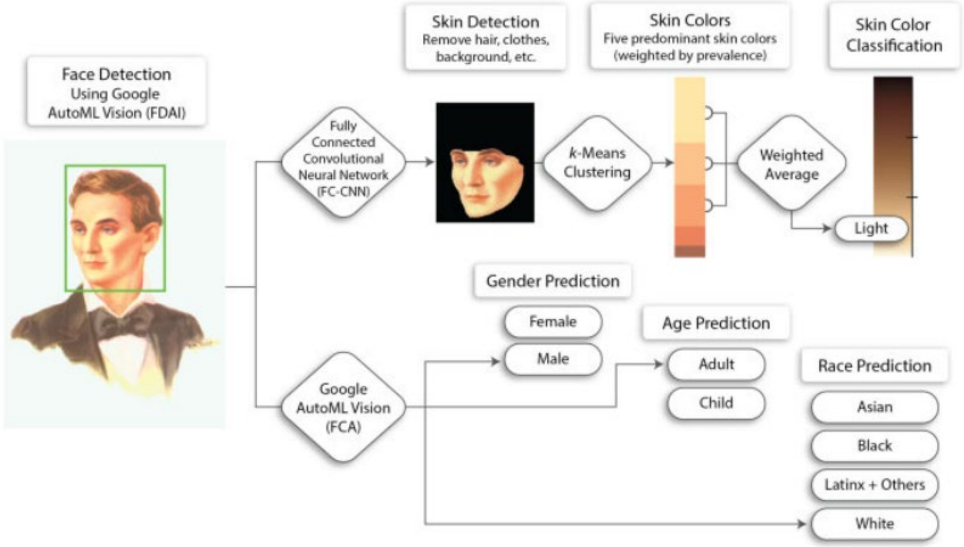
The Quarterly Journal of Economics, Volume 138, Issue 4, November 2023, Pages 2225–2285, https://doi.org/10.1093/qje/qjad028

Published: 31 August 2023 **Article history** ▾

PDF Split View Cite Permissions Share ▾

WHAT WE TEACH ABOUT RACE AND GENDER

(A)



Nontraditional datasets

(Almost) 200 Years of News-Based Economic Sentiment*

J. H. van Binsbergen[†] S. Bryzgalova[‡] M. Mukhopadhyay[§] V. Sharma[¶]

December 2023

Abstract

Using text from 200 million pages of 13,000 US local newspapers and machine learning methods, we construct a 170-year-long measure of economic sentiment at the country and state levels, that expands existing measures in both the time series (by more than a century) and the cross-section. Our measure predicts GDP (both nationally and locally), consumption, and employment growth, even after controlling for commonly-used predictors, as well as monetary policy decisions. Our measure is distinct from the information in expert forecasts and leads its consensus value. Interestingly, news coverage has become increasingly negative across all states in the past half-century.

Keywords: Business cycle, macroeconomic news, economic sentiment, monetary policy, textual analysis, machine learning, big data, neural networks
JEL codes: G1, G4, E2.

The Ghost in the Machine: Generating Beliefs with Large Language Models *

J. Leland Bybee

Yale University

First Draft: February 16, 2023
This Draft: November 15, 2023
Updated Regularly: Latest Version

Abstract

I introduce a methodology to generate economic expectations by applying large language models to historical news. Leveraging this methodology, I make three key contributions. (1) I show generated expectations closely match existing survey measures and capture many of the same deviations from full-information rational expectations. (2) I use my method to generate 120 years of economic expectations from which I construct a measure of economic sentiment capturing systematic errors in generated expectations. (3) I then employ this measure to investigate behavioral theories of bubbles. Using a sample of industry-level run-ups over the past 100 years, I find that an industry's exposure to economic sentiment is associated with a higher probability of a crash and lower future returns. Additionally, I find a higher degree of feedback between returns and sentiment during run-ups that crash, consistent with return extrapolation as a key mechanism behind bubbles.

First insight: AI can be used to *unlock* data

- There is more nontraditional data than tabular data.
- Problem: it's locked inside text, pictures, audio, etc.
  - We can't run OLS on text
- Solution: use LLMs to extract such data!

Example 1: Large amounts of public data

Regulations.gov
Your Voice in Federal Decision Making

Docket (FWS-HQ-NWRS-2022-0106) / Document

PR PROPOSED RULE

National Wildlife Refuge System: Biological Integrity, Diversity, and Er

Posted by the Fish and Wildlife Service on Feb 2, 2024

Comment | View More Documents 3 | View Related Comments 33.02K | Share

Document Details | Browse Posted Comments 33.02K

- 33 thousand comments!
- How do we make sense of them? Summarize them, classify them, etc.

You

FWS-HQ-NWRS-2022-0106-32972_att...
PDF

The attached document is a comment on a proposed rulemaking. Use this document to respond the following questions:

1. Who wrote the comment?
2. Did the author represent any interest group?
3. Is the comment in favor or against the proposed rule?
4. What is the argument used in favor or against this rule?
5. Does the comment proposes any alternative rules? Which ones?

ChatGPT

1. The comment was written by Karimah Schoenhut, Senior Staff Attorney for the Sierra Club.
2. Yes, the author represented the Sierra Club, which is the nation's oldest grassroots environmental organization, dedicated to the protection and preservation of the environment.
3. The comment is against the proposed regulations for the National Wildlife Refuge System regarding Biological Integrity, Diversity, and Environmental Health ("BIDEH") and the updated BIDEH policy. Sierra Club supports the recognition of the need to ban predator control of native predators on Refuge System lands and to provide direction for requiring mitigation of greenhouse gas emissions but urges the Service to strengthen the proposed regulations to better ensure fulfillment of the BIDEH mandate

Example 2: Extract historical data

You

The text pasted below is a snippet from a German text from 1929. It might contain employment information by year (number of employees). If avaiable, please extract it in a two-table column, where the first table is the year and the second is the number of employees:

Waldthausen, Essener Credit-Anstalt; Cöln u. Berlin: A. Schaaffh. B.-V.;
Aachen: Rhein.-Westf. Disconto - Ges.; Bexlip.: Deutsche Bank, S. Simonson.
Dividenden 1863—1913: 0, 0, 2%, 0, 3' 3, 4, 6, 20, 25, 40, 25, 10, 0, 0, 0, 0, 6, 3,
5, 6, 6, 6,17, 8, 15, 30, 80, 80, 40, 30, 35, 40, 50, 60, 65, 75, 50,"45, 35, 40, 35, 35, 45,
45, 24 (j. 12), 137s, 16, 16%, 22, 25%. — Cpnverj. 4 Jahre (K).

(Many lines later…)

Grundbesitz 1910-1913: 1374, 1403, 1388, 1413 ha.
Häuser 1910—1913 : 757, 805, 1144, 1351.
Oeffentliche Lasten 1909—1913 (Mill. dH): l-.ra. 1.76, l.73, 1-82, l-oa
Arbeiter 1911—1913 : 7114, 7535, 8358.
Gewinn 1913: Vortrag 61171, Betrieb u. Diverse 6 796 738, Gew. Ar
zus. 6 157 909 dH. — Ab: Oblig.-Zinsen 45 200, Abschreib. 1906
dH.

ChatGPT

The snippet from the German text mentions employee numbers for the years 1911 to 1913. Below is the requested two-column table with the years and the corresponding number of employees:

Year	Number of Employees
1911	7114
1912	7535
1913	8358

This extraction reflects the only years and employee numbers explicitly mentioned in the provided text.

"The Debt-Inflation Channel of the German Hyperinflation"
(w/ Brunnermeier, Luck, Verner, Zimmermann)

Example 3: Create historical data

SE **You**

The text pasted below is a snippet from a German text from 1919, referencing a firm activities over the last year. It might contain information on firm real investments, such as factories, warehouses, etc.. If available, please list any such investments in a three-column table, where the first column is the name in German, the second is the name in English, and the third is the amount invested, if available.

SNIPPET:

„Adler" Deutsche Portland-Cement-Fabrik, Act.-Ges.
Vorstand: Anton Piper (Berlin), Dr. phil. Heinr. Müller (Rüdersdorf).
Aufsichtsrat: Vors.: Justizrat Dr. Max Hirschei; sonstige Mitglieder: Bank-
Dir. Curt Sobemheim [Commerz- u. Diseonto-Bk.], Kommerz.-IL G. Schröter
[Deutsche Bk.], Justizrat Th. Marba [Nationalbk. f. D.j. — Fingetr.: 13./4. 1880.
Sitz in Berlin mit Zweigniederlass. in Berlin-Wilmersdorf, Nikolsburger
platz 6/7. — Die Ges. übernahm bei der Gründ. 198 Morgen 175 qR nahe der
Stadt Zossen Kalkwiesen u. Tonläger nebst der auf diesen Grundst. befind).
Cementfabrik u. Dampfziegelei für l 000 000 c/F; später noch 300 Morgen in der
Nachbarschaft der Fabrik gelegene Kalkwiesen u. Tonläger. — Am 23-/5. 1899
vom A.-R. Ankauf der Portland-Cement-Fabrik Rüdersdorf R. Guthmann Je
Jeserich einschl. der Aussenstände für 8 000000 kV beschlossen.
Besitz: Portland-Zement-Fabrik mit Kalkwerk in Zossen, 3 Porti.-Zemeut-
i'abr. u. Kalkwerk in Tasdorf b. Kalkberge-Rüdersdorf. Rüdersdorfer Besitz
i-a. 63 Morgen. Die Zementfabrik l besitzt 4 Drehöfen, die ebenso wie eine

ChatGPT

Based on the provided snippet from a 1919 German text regarding the "Adler" Deutsche Portland-Cement-Fabrik, Act.-Ges., here is a summary of the firm's real investments in a table format:

German Name	English Name	Amount Invested
Kalkwiesen u. Tonläger nahe der Stadt Zossen	Lime meadows and clay deposits near Zossen	1,000,000 Mark
Kalkwiesen u. Tonläger in der Nachbarschaft der Fabrik	Lime meadows and clay deposits in the vicinity of the factory	Not specified
Portland-Cement-Fabrik Rüdersdorf R. Guthmann Je Jeserich	Portland Cement Factory Rüdersdorf R. Guthmann Je Jeserich	8,000,000 Mark
Portland-Zement-Fabrik mit Kalkwerk in Zossen	Portland Cement Factory with lime plant in Zossen	Not specified

- Create quantifiable measures of real investment
- Not just extract preexisting ones

Example 4: Merge datasets

- We often need to match observations across multiple datasets

- But IDs might not match!
  - “Wheeling NB”
  - “Wheeling, N. of W. Va.”
  - “NB of West Virginia at Wheeling”
  - “Wheeling NB of WV”

- Soln.: ~~fuzzy matching~~ LLMs

SE **You**

Below there is a list of West Virginia National Banks from 1929. The first column contains the state, the second contains the charter number, and the third contains the bank name. Please return the charter number and bank name of the bank that most closely matches the alternative bank name "Wheeling, N. of W. Va". If no bank is a close match, please return "0" as the charter number.

WV	180	FNB of Parkersburg
WV	360	FNB of Wheeling
WV	864	Second NB of Parkersburg
WV	961	FNB of Fairmont
WV	1343	Merchants NB of West Virginia at Wheeling
WV	1387	FNB of Wellsburg
WV	1424	NB of West Virginia at Wheeling
WV	1427	Parkersburg NB
WV	1502	Merchants NB of West Virginia at Morgantown
WV	1504	Merchants NB of West Virginia at Point Pleasant
WV	1524	NB of Martinsburg
WV	1530	Merchants NB of West Virginia at Clarksburg

ChatGPT

The bank that most closely matches the alternative bank name "Wheeling, N. of W. Va" is:

Charter Number: 1424

Bank Name: NB of West Virginia at Wheeling

“Failing Banks”
(w/ Luck, Verner)

Second insight: AI works *at scale*

- A good intern can do anything that an LLM can do
- But with AI we can do so at scale
- What headcount would an academic researcher need to:
  - Search 40,000 pages of old, hard-to-read German
  - Go through 10,000 PDFs to extract information
- Now imagine as a PhD student!

Act 2:

Google Colab demo at:

http://scorreia.com/llm

Background - Tokenization

- To process text, images, audio, etc., we must first convert them into numeric vectors (of varying length)
- At its most trivial, tokenization is:

```python
text = 'Hello world!'
vector = [ord(x) for x in text]
vector
[72, 101, 108, 108, 111, 32, 119, 111, 114, 108, 100,
33]
```

Background - Tokenization

Input text:

In 2023 we saw a virtually unprecedented decline in inflation without an associated recession. In a Phillips curve framework, 2023 saw one of the largest ever drops in inflation in a year, yet unemployment remained below the natural rate.

Tokenized text:

In 2023 we saw a virtually unprecedented decline in inflation without an associated recession. In a Phillips curve framework, 2023 saw one of the largest ever drops in inflation in a year, yet unemployment remained below the natural rate.

Integer representation

[644, 220, 2366, 18, 584, 5602, 264, 21907, 31069, 18174, 304, 25544, 2085, 459, 5938, 39621, 13, 763, 264, 37514, 16029, 12914, 11, 220, 2366, 18, 5602, 832, 315, 279, 7928, 3596, 21701, 304, 25544, 304, 264, 1060, 11, 3686, 26690, 14958, 3770, 279, 5933, 4478, 627]

Background - Tokenization

DEMO

Input text:

In 2023 we saw a virtually unprecedented decline in inflation without an associated recession. In a Phillips curve framework, 2023 saw one of the largest ever drops in inflation in a year, yet unemployment remained below the natural rate.

Tokenized text:

In 2023 we saw a virtually unprecedented decline in inflation without an associated recession. In a Phillips curve framework, 2023 saw one of the largest ever drops in inflation in a year, yet unemployment remained below the natural rate.

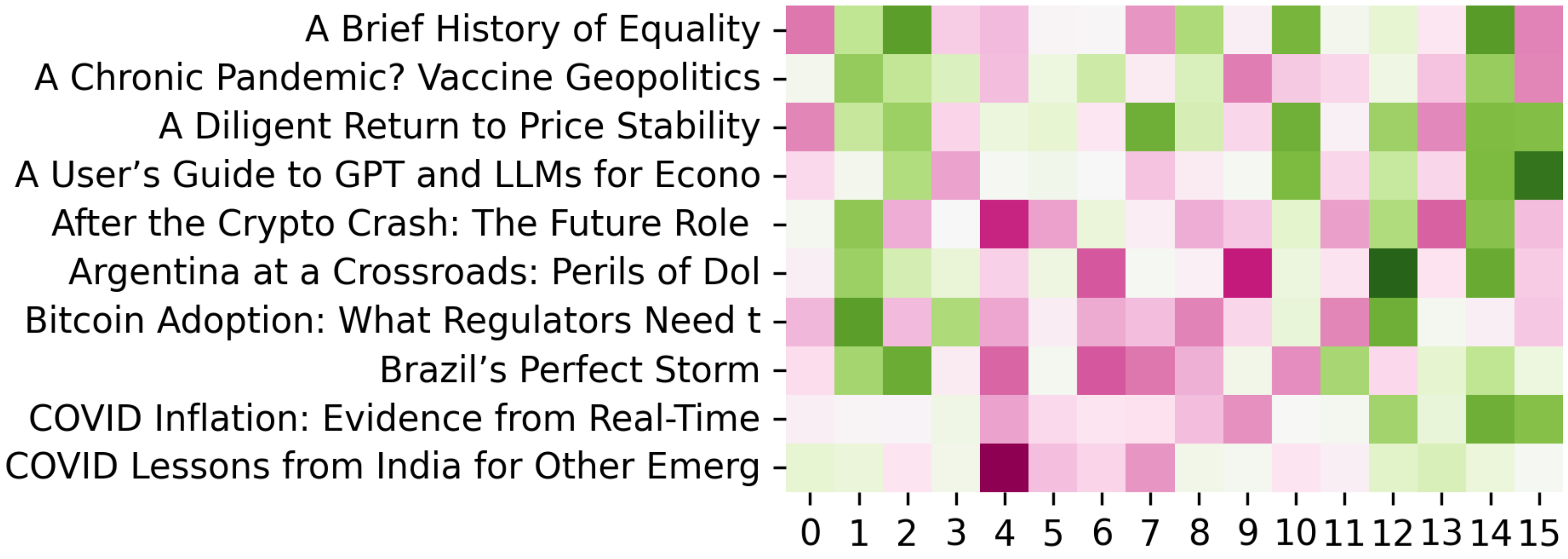
Integer representation

[644, 220, 2366, 18, 584, 5602, 264, 21907, 31069, 18174, 304, 25544, 2085, 459, 5938, 39621, 13, 763, 264, 37514, 16029, 12914, 11, 220, 2366, 18, 5602, 832, 315, 279, 7928, 3596, 21701, 304, 25544, 304, 264, 1060, 11, 3686, 26690, 14958, 3770, 279, 5933, 4478, 627]

Problem: Varying length; based on characters and not a deeper meaning

Background - Embeddings

- Low-dimensional semantic representations of tokens
- Embedding space = Latent space (think PCA) encoding an underlying meaning

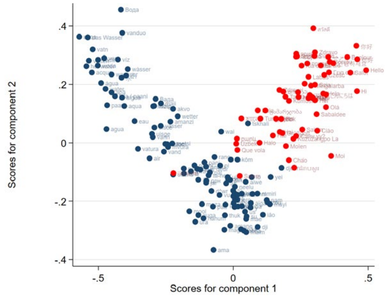


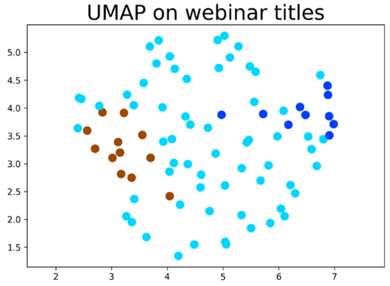
Background - Embeddings

Even without LLMs, embeddings can do many things!

- Linear algebra:
  - Add, subtract, average them (word2vec's "king+woman-man = queen")
  - Compute distances across texts (cosine similarity) DEMO

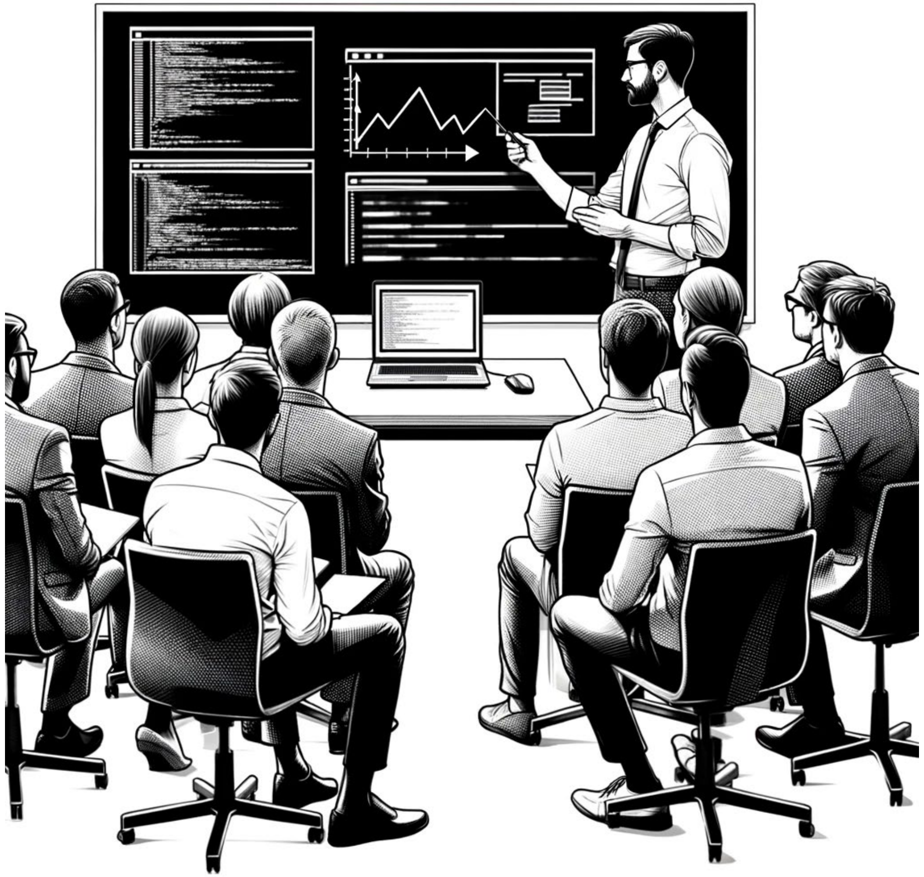
- Classification:
  - PCA, k-means clustering, t-SNE, UMAP





- Econometrics:
  - Use embeddings directly as RHS variables in lasso, elastic nets, random forests, etc. (Amazon review scores, stock returns around earning calls)

(Live demo)



Background – LLMs and GPTs

- Extremely sophisticated version of the auto-predict feature of messaging and email apps
  - GPT4 has 1.8 trillion parameters across 120 layers
  - Trained on 13 trillion tokens (~syllables) including books, papers, movie subtitles, etc.
  - Llama 2 has 70 billion parameters and was trained on 2 trillion tokens
- But have a crucial limitation…

Problem: LLMs don't know *your* data

Prompt →

SE **You**

How many national banks mergers occurred in the year 2022? Reply with just a single number.

ChatGPT

162

SE **You**

How many national banks mergers occurred in the year 1899? Reply with just the number of mergers, nothing else.

ChatGPT

I'll need to conduct some research to find the exact number of national bank mergers that occurred in the year 189 . [>-]

I couldn't fin ional bank mergers that The information ocuses on broader tren e changes affecting ba cific annual counts, especially for the late 19th and early 20th

Solution: add context to the LLM

Prompt:

You

You have been tasked to extract information from an old German investment handbook from the 1910s. You will find information corresponding to a single firm after the word "CONTEXT:" at the end of this message.

Your objective is to extract information on real assets purchased by this firm, including factories, warehouses, machinery, etc. Provide the solution in a table with four columns:

1. german_name: Name of the asset in German
2. english_name: Name of the asset in English
3. year: The year when this asset purchase took place, if available
4. number: Number of assets owned

Further, follow these instructions:

1. Please only answer the question using the provided context.
2. Note that the text is based on an OCRed scan, so it might contain typos typical of OCRed documents.
3. Exclude from the list of assets the stated purpose of the firm, often prefixed by the "Zweck:" label (or a similar label).

+ Context

4. Hint: the list of assets is usually prefixed by a label such as Gegenwärtiger Besitz, Besitz, Besitzum, Besitzstand (and variations).

CONTEXT:

(Calbe a. S.), Rentier Sigism. Sternberg (Charlottenburg).
Errichtet: 11./3.
1873. Ges. - Kasse in Völpke b. Magdeburg. - Besitztum: Der Grubenbesitz
umfasst 6 zum grössten Teil bereits abgebaute verliehene Braunkohlenfelder.
in den Gemarkungen Völpke, Offleben u. Barneberg mit einem noch gewinn-
baren Kohlenquantum von rund 100 000 000 hl. Hingegen ist die Kohlenab-
baugerechtigkeit über rd. 660 Morgen im Bitterfelder Braunkohlen - Revier in
der Gemarkung Roitzsch mit einem nach rechnungsmässiger Schätzung sich
ergebenden Kohleninhalt von rd. 320 000 000 hl gewinnbarer
Kohle unverritzt

= Answer

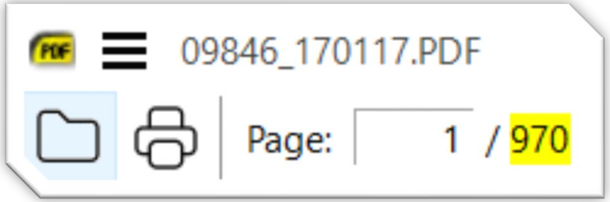
ChatGPT

Based on the provided context, I have extracted the information regarding real assets purchased by the firm. Here's the table:

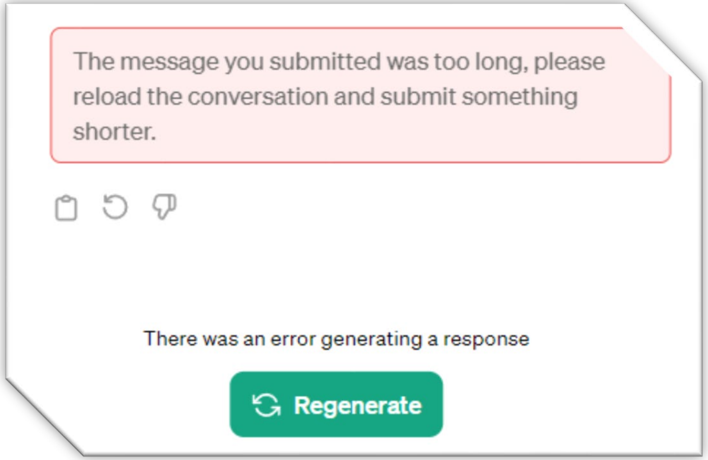
german_name	english_name	year	number
Grubenfelder	Mining Fields	-	6
Braunkohlenfelder	Brown Coal Fields	-	6
Braunkohlen-Bergwerk	Brown Coal Mine	-	1
Brikettfabrik	Briquette Factory	1899	1
Drahtseilbahn	Wire Rope Railway	-	1
Dampftelleröfen	Steam Plate Ovens	-	5
Dampfkessel	Steam Boilers	-	10
Kühlanlage	Cooling System	-	1
Dampf- u. elektr. Maschinen	Steam and Electric Machines	-	multiple
Dampflokomotiven	Steam Locomotives	-	2
Beamten- u. Arb.-Wohnhäuser	Officer and Employee Residences	-	16

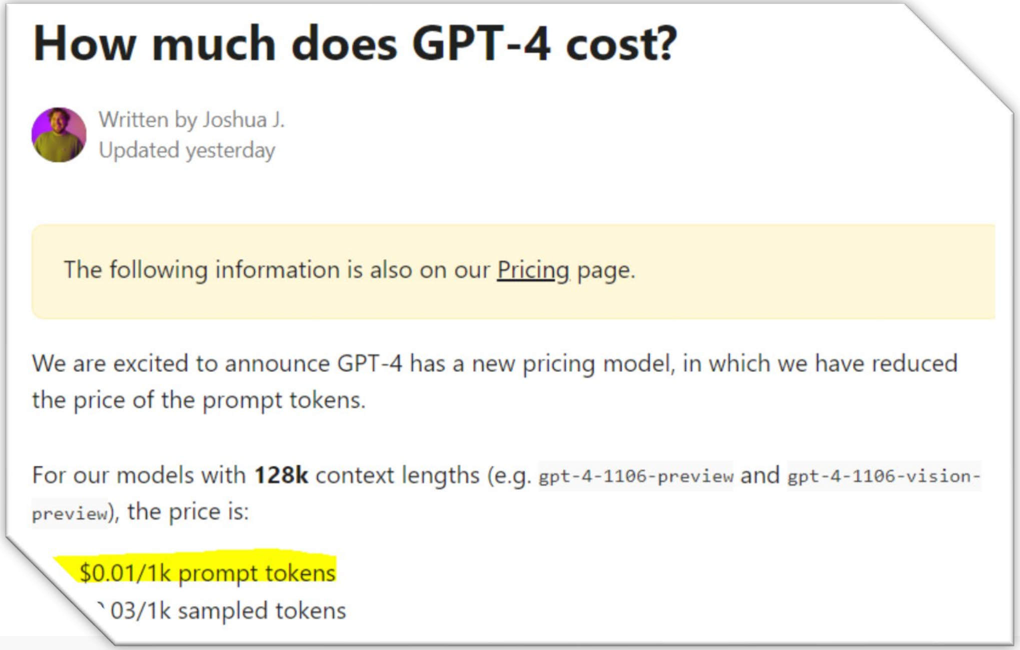
Caveat: context should be small

Input:

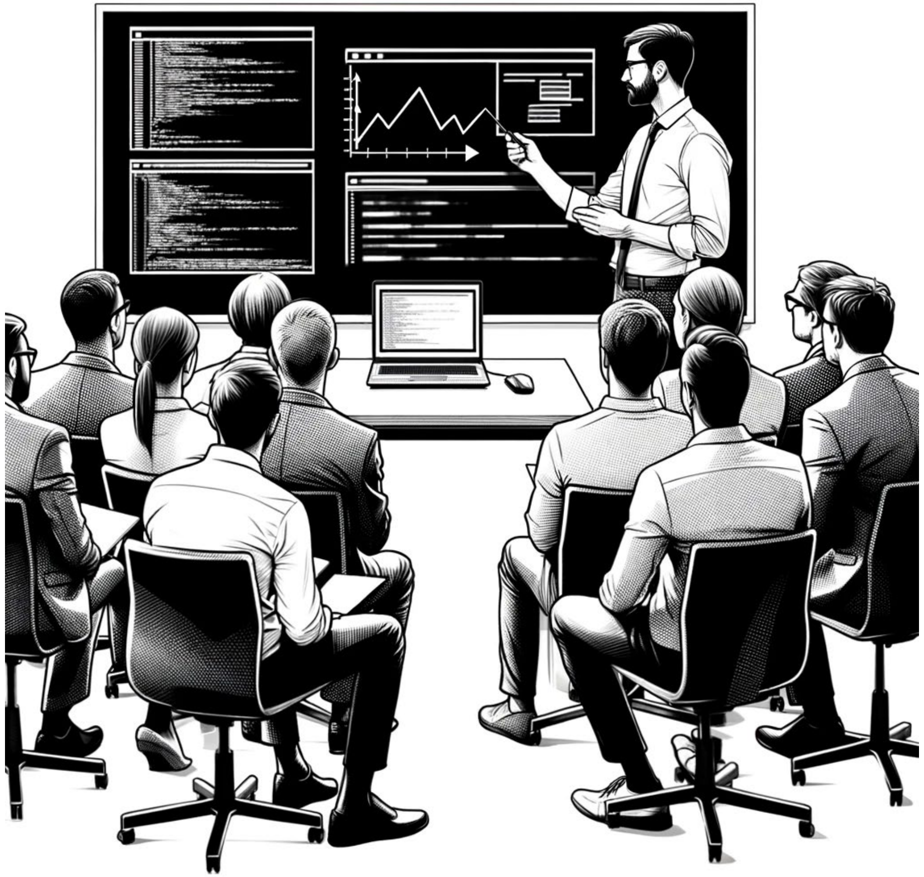


Output:





(Live demo)



Putting it all together: Retrieval-Augmented Generation (RAG)

1. Read documents and split them into chunks
2. Build chunk embeddings
3. Given a search query, build the query embedding
4. Select chunks closest to query in embedding space
5. Build prompt that includes selected chunks

(Beware: for research projects, frameworks are not as useful as expected, and could even be detrimental)

Act 3:

Trivial RAGs are easy; good RAGs are *hard*

Leonie @helloiamleonie · 20h

Building a prototype of a **RAG** application is a piece of cake - but making it production-ready is a real challenge.

Mayo Oshin
@mayowaoshin

How to Analyze Tables In Large Financial Reports Using GPT-4 (w/ @jerryjliu0)

Most corporate docs contain a mix of text and tables. But if you use simple RAG split and chunk methods, the AI model will likely hallucinate due to embedding split tables.

Trivial RAGs are easy; good RAGs are *hard*

RAG pipelines can get quite complex

1. PDF parsers
2. Chunking strategies (fixed-size vs semantic)
3. Choice of embedder; choice of vector database
4. Search method (approximate kNN), metadata filtering; reranking
5. Optimal number of chunks and context window
6. Prompt engineering; temperature
7. Etc. etc. etc.

New things often break

Feb 14, 2024

Elevated error rate impacting Ch

Resolved - This incident has been resolve
Feb 14, 12:40 PST

Monitoring - A fix has been implemented
Feb 14, 10:38 PST

Investigating - We are continuing to inves
Feb 14, 10:20 PST

Elevated error rate impacting AP

Resolved - This incident has been resolve
Feb 14, 12:40 PST

Monitoring - A fix has been implemented
Feb 14, 10:38 PST

Update - We are continuing to investigate
Feb 14, 07:08 PST

Investigating - We are currently investiga
Feb 14, 07:06 PST

Feb 13, 2024

Elevated errors across

Resolved - This incident has
Feb 13, 15:38 PST

Monitoring - A fix has been
Feb 13, 13:46 PST

Identified - We have identifi
Feb 13, 12:38 PST

Investigating - We are curre
Feb 13, 12:11 PST

Feb 12, 2024

Elevated errors on GP

Resolved - This incident has
Feb 12, 10:39 PST

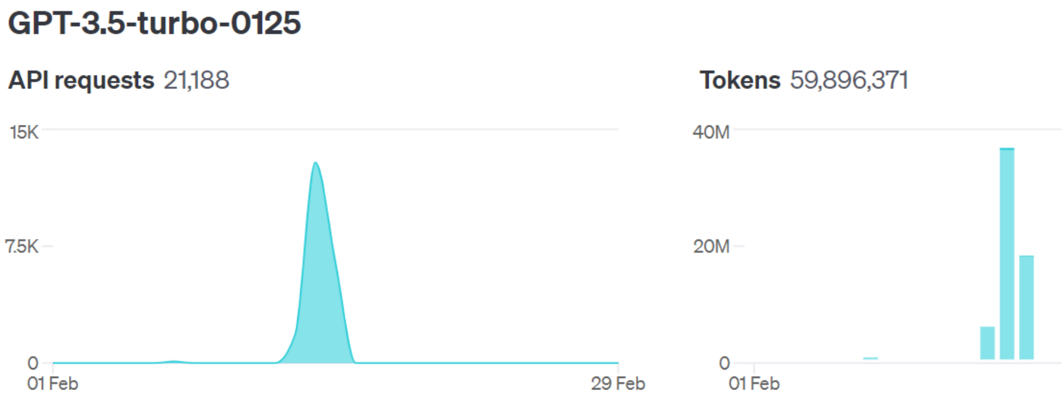
Update - We are continuing
Feb 12, 10:19 PST

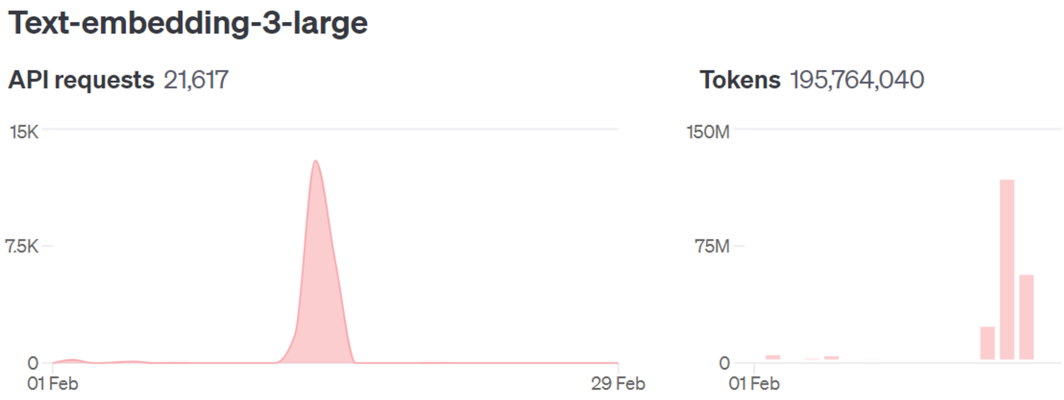
Monitoring - A fix has been
Feb 12, 10:19 PST

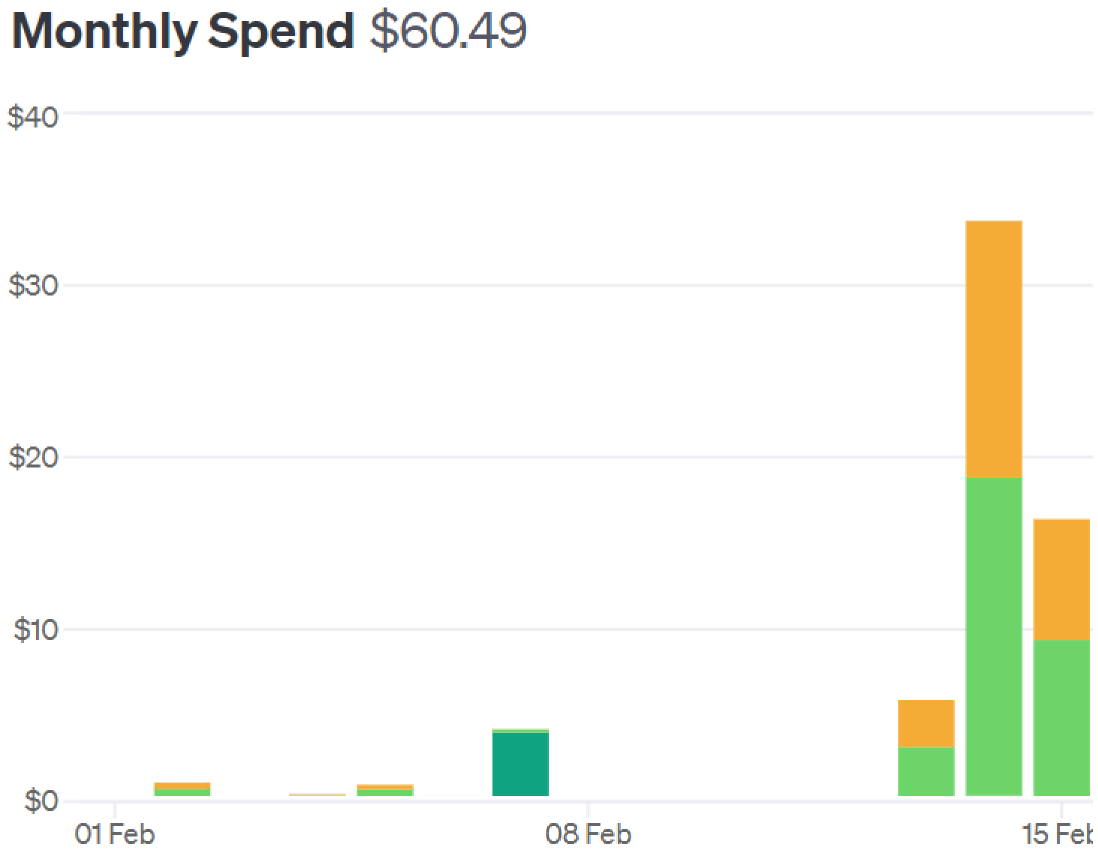
Update - We are continuing
Feb 12, 09:19 PST

Identified - We are currently
Feb 12, 09:19 PST

Shiny tools cost (a little) money







Can we trust our results?

- Should probably build some ground truth
  - And test against it
  - But don’t over fit!
- How to ensure reproducibility
  - Random seeds?
- How to limit hallucinations
  - Lower “model temperature”?
  - Use multiple independent models?

Summary

1. LLMs allow us to unlock data for research

2. Could be done before, but at a great cost

    - **Bold claim**: These tools democratize access to these data
    - Don't need an army of interns or substantial research funds
    - Still need humans for validation though!

3. Python + LLMs = Incredibly powerful combo
    - No need to be a professional coder or Python expert

Thank you!

- Colab link:
  - http://scorreia.com/llm
- Further reading:
  - Andrej Karpathy's brilliant course: https://karpathy.ai/zero-to-hero.html
  - Broader course by fast.ai https://course.fast.ai/
  - Ethan Mollick's guides to prompting: https://www.oneusefulthing.org/p/working-with-ai-two-paths-to-prompting
  - Simon Willinson's "Embeddings: what they are and why they matter" https://simonwillison.net/2023/Oct/23/embeddings/