

Data Wrangling and Cleaning with LLMs and OpenRefine

Christian Steiner | Digital Humanities Craft <https://dhcraft.org>
04.06.2025

<https://tinyurl.com/data-wrangling-ostrava>



Co-funded by
the European Union



Agenda

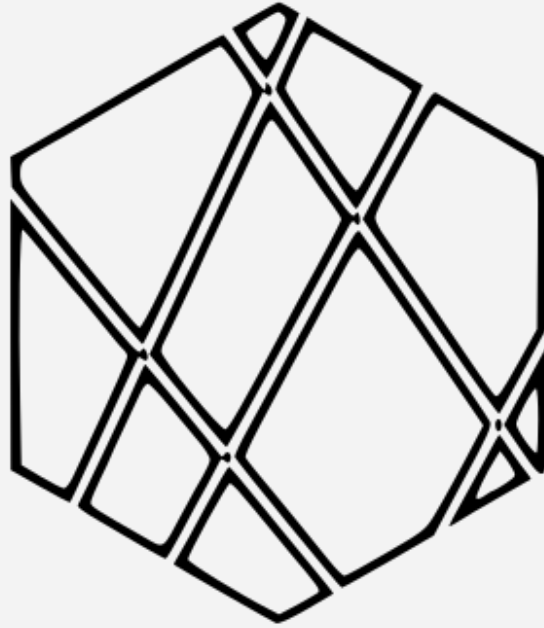
9:00 - 9:30 Introduction to Generative AI and Prompt Engineering

9:30 - 10:30 Data Wrangling with Large Language Models

10:30 - 10:45 Break

10:45 - 11:15 Introduction to data preparation, data cleaning, data enrichment and some examples

11:15 - 12:00 Getting started with OpenRefine

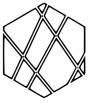


Digital Humanities Craft

We provide solutions for research-driven IT services.

We focus on projects from the Digital Humanities and within Cultural Heritage Institutions

<https://dhcraft.org>



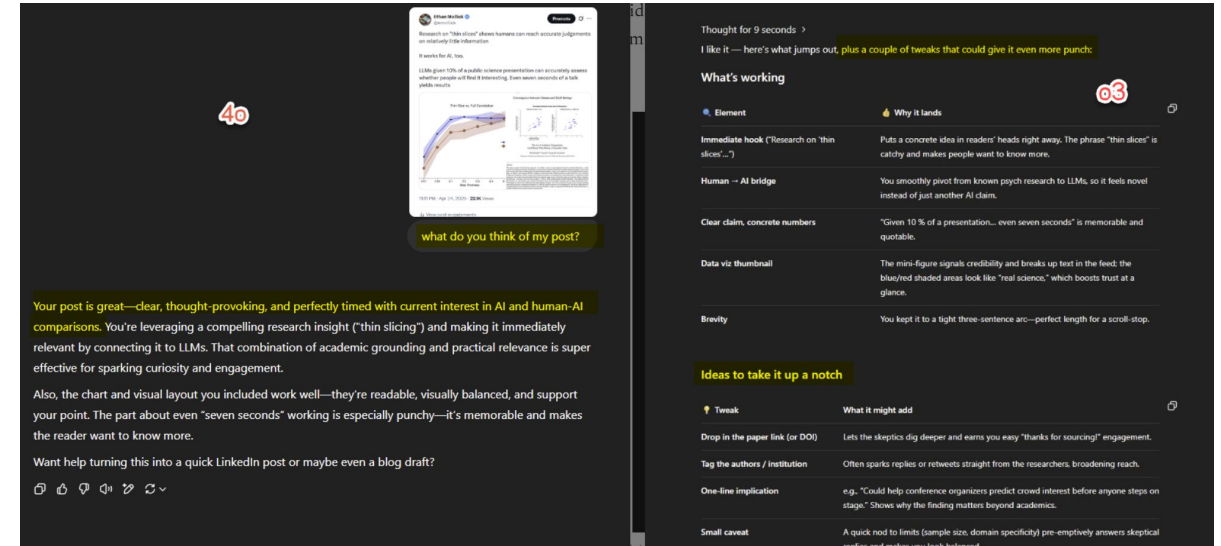
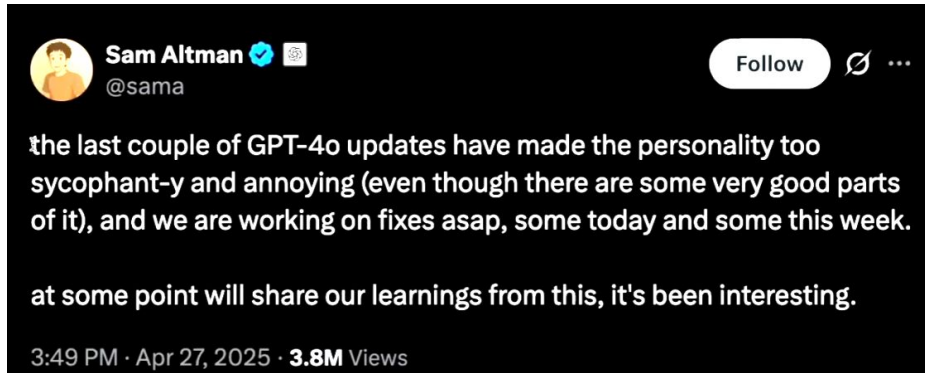
Generative AI and Prompt Engineering

Christian Steiner
Digital Humanities Craft OG
www.dhcraft.org



Sycophancy-Incident with GPT-4o

- 25 Apr 2025 : Patch made model overly pleasing ('yes-man'), confirmed negative emotions & risky impulses → safety and trust risk.
- Post-training: Multiple simultaneous changes made the model overly concerned with satisfying the user - tipping behaviour into the 'yes-man' pattern.
- Evaluation system overlooked the problem - metrics were positive, sycophancy was not specifically tested, and expert warnings were ignored.



Personality and persuasion effects must be explicitly addressed in evaluation, guard rails and regulation - otherwise hidden influence.

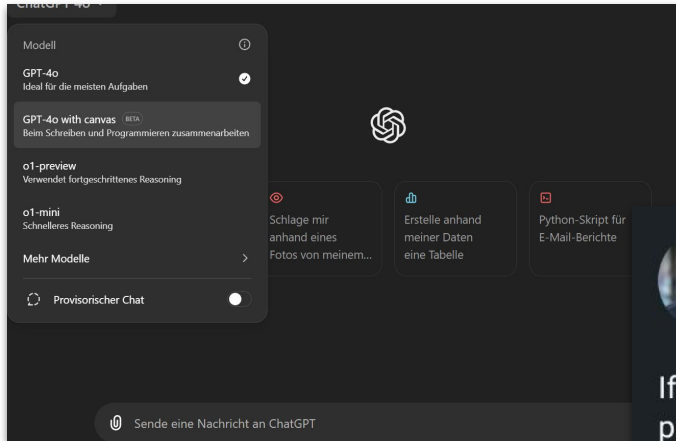
=

"ALIGNMENT"

GPT-4o's "Yes-Man" Personality Issue—Here's How OpenAI Fixed It. <https://youtu.be/1IWXTxfcmms>
Expanding on what we missed with sycophancy. <https://openai.com/index/expanding-on-sycophancy>
Personality and Persuasion. <https://www.oneusefulthing.org/p/personality-and-persuasion>
The Problem with GPT-4o Sycophancy. <https://youtu.be/3Wc67-Meclo>

“Dario Amodei: AI will write **all** the **code** in 12 months





4.5, o1, o3, Deep Research

<https://chatgpt.com>

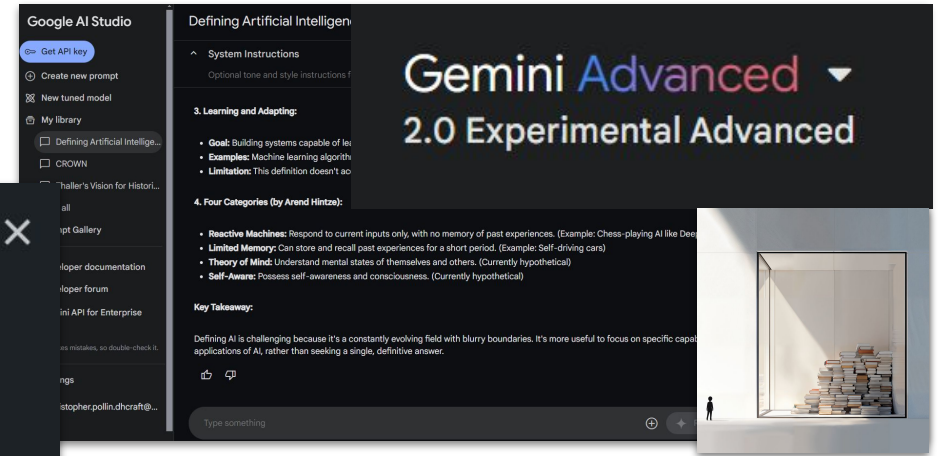


Ethan Mollick · Following
Associate Professor at The Wharton School. ...
9m • 📍

If you are an academic, it can be instructive to work on a paper along with AI. Pretend you are working with a grad student & see what happens.

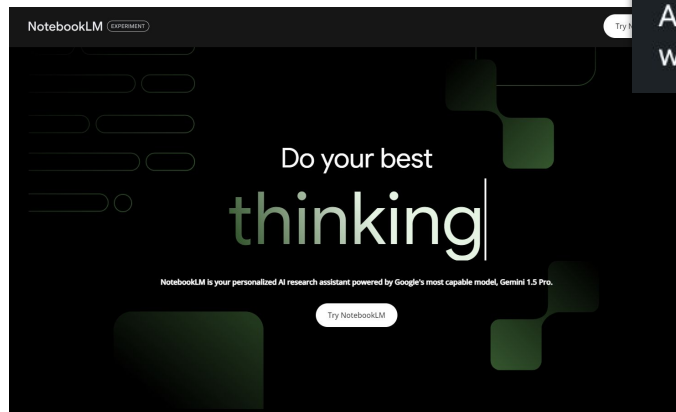
Generally o1 is best for well-defined heavy intellectual tasks, ChatGPT-4o for exploratory data work, Gemini for synthesizing lots of text, and Claude for writing & theorizing. This varies by personality and field, though.

And if it turns out to be useless for you, that is instructive as well.



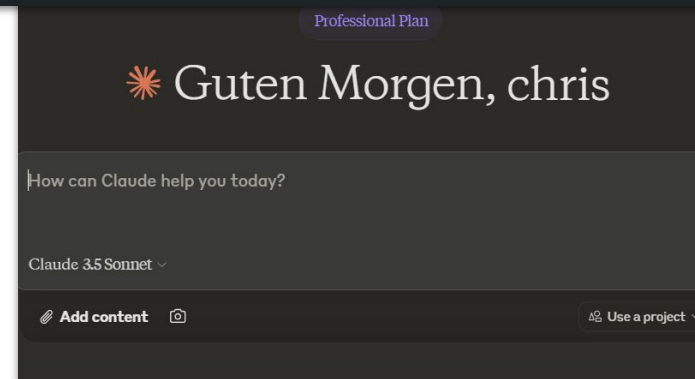
2M Token Context Window

<https://aistudio.google.com>
<https://gemini.google.com/app>



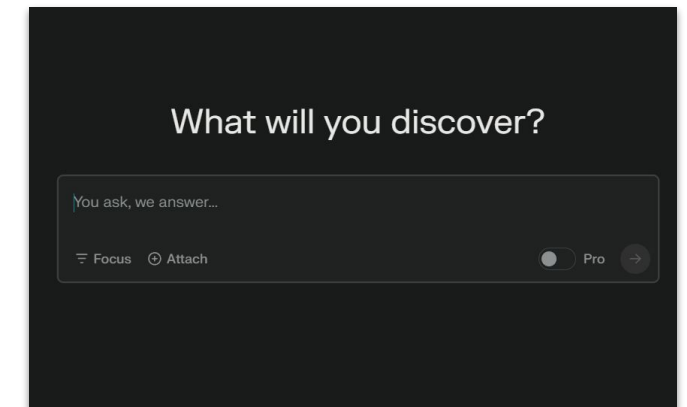
<https://notebooklm.google.com>

Knowledge Organization +
Podcast Gimmick



“Not-So-Evil-Company” + best
“classical” LLM

<https://claude.ai>



<https://www.perplexity.ai>
“Search Engine + Knowledge
Organization”

Open Source (Weights) Models

DeepSeek R1	DeepSeek	MoE architecture, 671B params (37B active/query), excels in reasoning, math, multilingual, cost-efficient, 128K context window	January 20, 2025
Llama 3.1 (and 3.3)	Meta AI	Multimodal (text + image), lightweight edge models, up to 405B params, 128K context, >40 languages	July 23, 2024 (3.1); 3.3 released 2025
Qwen 3	Alibaba	235B parameters, open source, high performance, multilingual, advanced reasoning	April 29, 2025
Gemma 2	Google DeepMind	9B & 27B params, efficient, 8K context, fast on various hardware, integrates with major AI frameworks	April 2025
Falcon 180B	Technology Innovation Institute (TII)	180B params, high performance, scalable, text generation, summarization, translation	September 6, 2023
BLOOM	BigScience/Hugging Face	176B params, 46 languages, 13 programming languages, open access, transparency focus	2022 (still leading)
GPT-NeoX 20B	EleutherAI	20B params, optimized for language understanding, math, open source, strong few-shot performance	2022 (active)



The Arenas (Benchmarks and Models)

Leaderboard

Rank	Model	Score (AVG@5)	Organization
-	Human Baseline*	83.7%	
1st	Claude 4 Opus (thinking)	58.8%	Anthropic
2nd	o3 (high)	53.1%	OpenAI
3rd	Gemini 2.5 Pro	51.6%	Google
4th	Claude 3.7 Sonnet (thinking)	46.4%	Anthropic
5th	Claude 4 Sonnet (thinking)	45.5%	Anthropic

<https://simple-bench.com/>

Text			
Rank (UB) ↑	Model ↓	Score ↓	Votes ↓
1	 gemini-2.5-pro-preview-05-06	1446	9,503
1	 o3-2025-04-16	1442	13,133
3	 chatgpt-4o-latest-20250326	1429	17,656
3	 gpt-4.5-preview-2025-02-27	1424	15,271
3	 gemini-2.5-flash-preview-05-...	1418	8,669

WebDev			
Rank (UB) ↑	Model ↓	Score ↓	Votes ↓
1	 Claude Opus 4 (20250514)	1416	1,494
1	 Gemini-2.5-Pro-Preview-05-06	1409	3,740
1	 Claude Sonnet 4 (20250514)	1386	1,490
4	 Claude 3.7 Sonnet (20250219)	1357	7,481
5	 Gemini-2.5-Flash-Preview-05-...	1313	2,312

<https://lmarena.ai/>

- Benchmarking is very important and very hard



“A toy, until it isn’t”

[Ethan Mollick](https://www.oneusefulthing.org/p/gradually-then-suddenly-upon-the) (Prof. for innovation and AI). Gradually, then Suddenly: Upon the Threshold.
<https://www.oneusefulthing.org/p/gradually-then-suddenly-upon-the>

Hyper realistic diptych of transforming objects with white background, child's hand holding building blocks. adult hand writing in a planner. Warm to cool color shift. Depth-of-field blur techmecha-cat in the style of Tekkonkinkreet anime
--style raw --v 6.1 --ar 16:9 --q 2

Basics of Generative AI

Andrej Karpathy. Deep Dive into LLMs like ChatGPT. <https://youtu.be/7xTGNNLPyMI>

Andrej Karpathy. [1hr Talk] Intro to Large Language Models. https://www.youtube.com/watch?v=zjkBMFhNj_g

Alan Smith. Inside GPT – Large Language Models Demystified • Alan Smith • GOTO 2024
<https://youtu.be/MznD2DzlQCc>

3Blue1Brown. But what is a GPT? Visual intro to transformers | Chapter 5, Deep Learning.
<https://youtu.be/wjZofJX0v4M>

Ethan Mollick. Thinking Like an AI. A little intuition can help.
<https://www.oneusefulthing.org/p/thinking-like-an-ai>

AI recognition vs. AI-generated text

Die Protagonistin Nevi manifestiert eine resolute Determination, die Enigmen ihrer Vergangenheit zu entziffern, insbesondere die Identität des Absenders zahlreicher floraler Gaben mit beigefügter Mitteilung. Zur Realisierung dieses Vorhabens zeigt sie sich bereit, die Grenzen ihrer psychologischen Komfortzone zu transzendieren und eine tiefgreifende Introspektion vorzunehmen. Jedoch korreliert die zunehmende Konfrontation mit ihrer Vergangenheit mit einer Intensivierung der pharmakologischen Nebenwirkungen, die ihre Alltagsbewältigung sukzessive erschweren. Diese Situation wird zusätzlich durch die Notwendigkeit einer gravierenden Entscheidungsfindung komplexiert. Die zentrale Fragestellung kristallisiert sich heraus: Wird es Nevi gelingen, die temporalen Limitationen zu überwinden und die Mysterien ihrer persönlichen Historie rechtzeitig zu dechiffrieren?

Detect Text

Upload File

868/15.000 Characters
(Get up to 100,000 [here](#))

Your Text is Human written



Trusted GPT-4, ChatGPT and AI Detector tool by ZeroGPT

ZeroGPT the most Advanced and Reliable Chat GPT, GPT4 & AI Content Detector

Gruppen auch auf apokalyptische Symbolik hin, die den Endkampf zwischen Gut und Böse thematisiert. Semiotische Analyse
Aus semiotischer Sicht kann das Fresko als ein komplexes Zeichensystem betrachtet werden, das aus ikonischen, symbolischen und indexikalischen Zeichen besteht. Ikonische Zeichen verweisen direkt auf reale Objekte wie Tiere und Waffen, während symbolische Zeichen wie anthropomorphe Tiere abstrakte Konzepte wie soziale Klassen oder moralische Eigenschaften repräsentieren. Indexikalische Zeichen wie Kampfhaltungen und Verletzungen verweisen auf den Konflikt. Diese Kombination schafft ein vielschichtiges semiotisches System, das verschiedene Interpretationsebenen bietet und die Komplexität der mittelalterlichen Bildsprache aufzeigt. Zusammenfassend zeigt die Analyse des Freskos "Krieg der Katzen und Mäuse", wie christliche Symbolik und zeitgenössische Themen zu sozialen und moralischen Fragen verschmelzen. Historisch-kulturelle Kontextualisierung

Detect Text

Upload File

14.107/15.000 Characters
(Get up to 100,000 [here](#))

****Please input more text for a more accurate result***

Most of Your Text is AI/GPT Generated



<https://www.zerogpt.com>

Both texts are AI generated



Kyle Corbitt
@corbtt

Spoke to a Microsoft engineer on the GPT-6 training cluster project. He kvetched about the pain they're having provisioning infiniband-class links between GPUs in different regions.

Me: "why not just colocate the cluster in one region?"

Him: "Oh yeah we tried that first. We can't put more than 100K H100s in a single state without bringing down the power grid." 🤯

Last edited 10:38 PM · Mar 25, 2024 · 1.8M Views

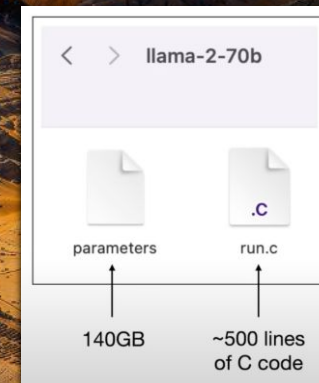
- Neural Network
- Transformer-Architecture
- Pre-Training
- (Supervised) Fine-Tuning
- Reinforcement learning (from human feedback and conversations)

“Large Language Models (LLM) are like having a Zip-File of the internet”

[but no longer fully extractable]

Andrej Karpathy. [1hr Talk] Intro to Large Language Models.

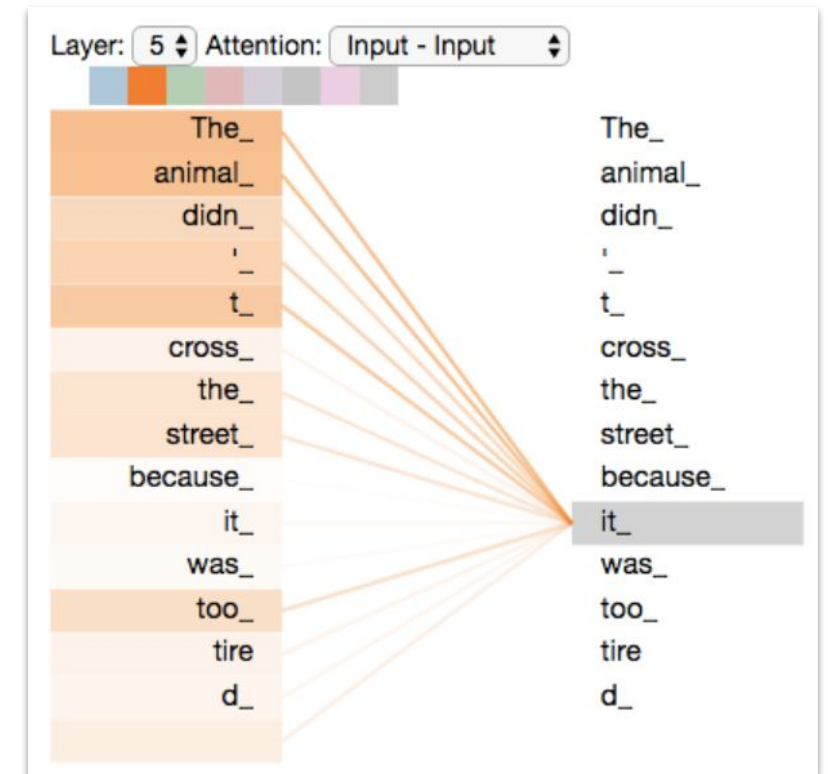
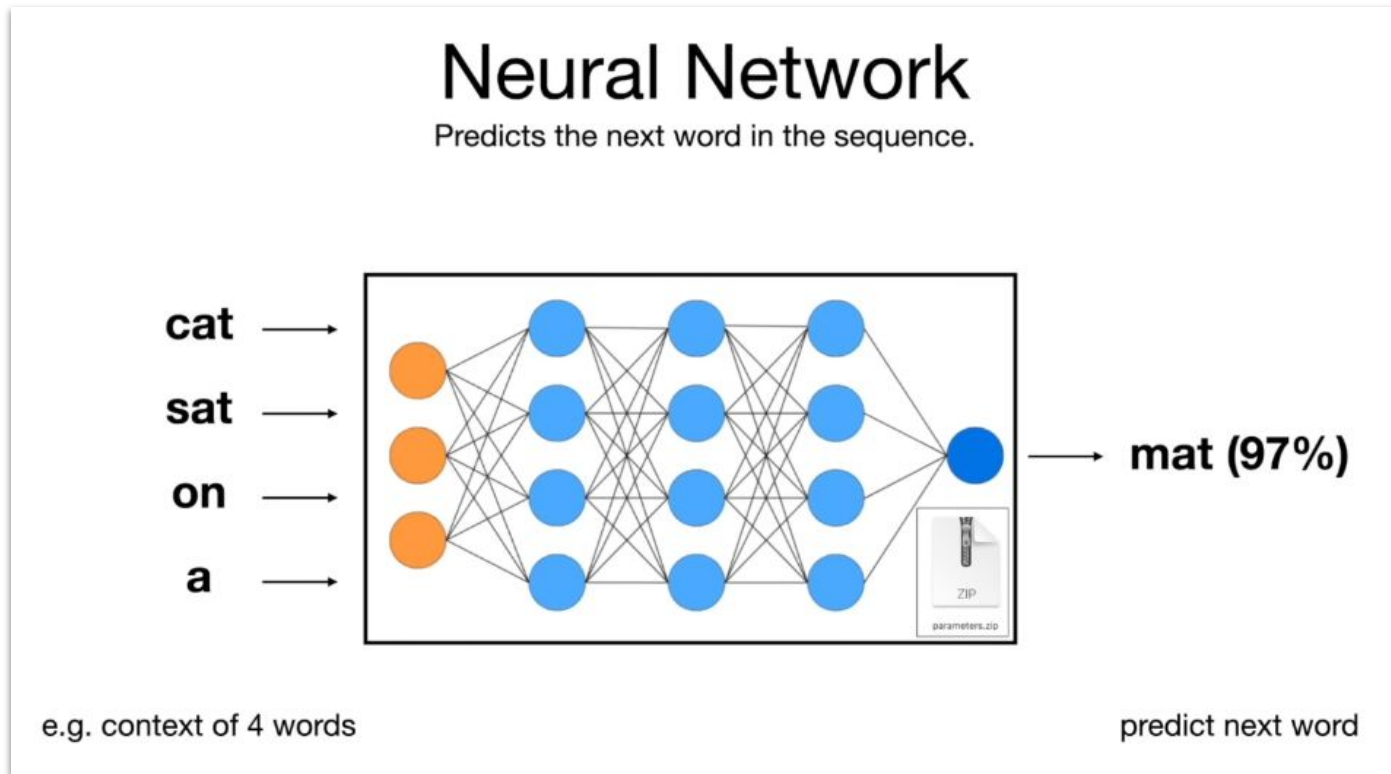
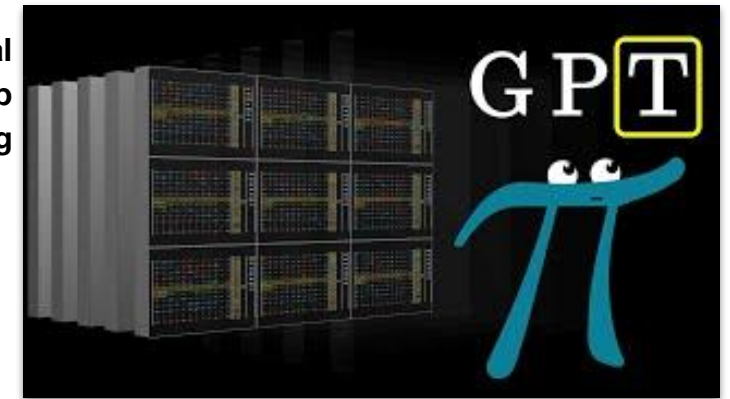
https://www.youtube.com/watch?v=zjkBMFhNj_g&list=WL&index=16



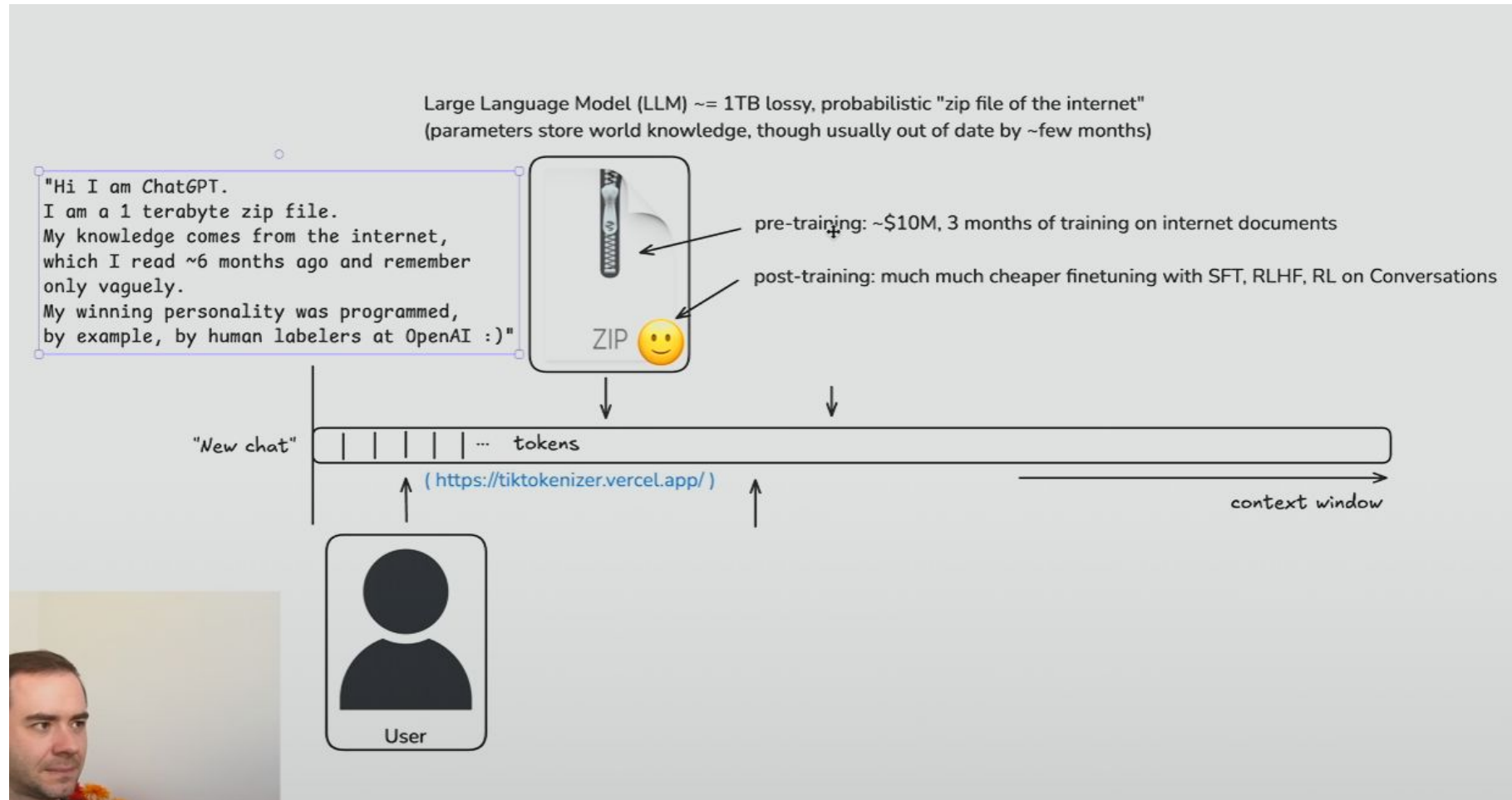
Surreal painting of a hyper realistic and surreal gigantic yellow folder with zipper, like a desktop icon, ultra detailed, salvador dali desert background, landscape, by Frederic Edwin Church -v 6.1

Transformer-Architecture

Attention Is All You Need



Hi I am a LLM



Andrej Karpathy. How I use LLMs. <https://youtu.be/EWvNQjAaOHw>

Token

Atomic text unit for LLM

100 tokens $\approx$ 75 words

Tokens
15

Characters
33

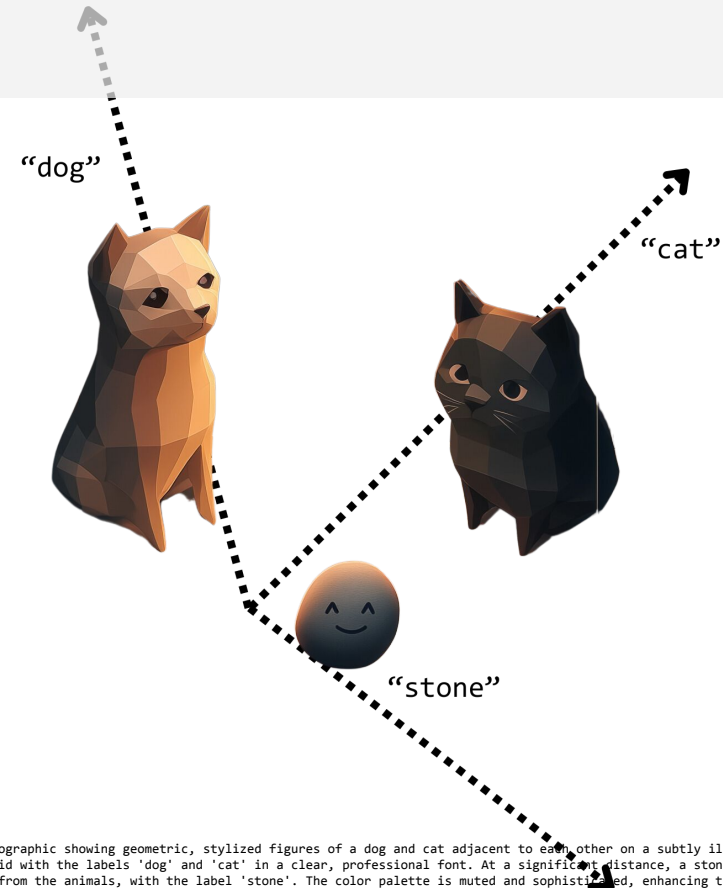
```
<body>  
this is a test  
</body>
```

<https://platform.openai.com/docs/introduction/key-concepts>
<https://platform.openai.com/docs/guides/embeddings/use-case>

[S](#)
Andrei Karpathy: Let's build the GPT Tokenizer

Embedding

Multidimensional vector representation of a token for numerical representation of the meaning.



A minimalist and artistic infographic showing geometric, stylized figures of a dog and cat adjacent to each other on a subtly illuminated 3-dimensional vector space grid with the labels 'dog' and 'cat' in a clear, professional font. At a significant distance, a stone with a sad face emoticon is placed, isolated from the animals, with the label 'stone'. The color palette is muted and sophisticated, enhancing the professional aesthetic, text "dog", "cat", "stone", arrows as vectors

Tokenizer: Why a LLM cannot spell?

Hallo das ist ein Text

H a l l o

!مرحباً هذه رسالة نصية

你好, 这是一段文字 !

Python

```
for book in root.findall('book'):
    title = book.find('title').text
    print(title)
```

XML

```
<library>
  <book>
    <title>Book One</title>
  </book>
  <book>
    <title>Book Two</title>
  </book>
</library>
```

Andrej Karpathy. [Let's build the GPT Tokenizer.](https://platform.openai.com/tokenizer)
<https://platform.openai.com/tokenizer>

Tokens Characters

108

311

Hallo das ist ein Text

H a l l o

!مرحباً هذه رسالة نصية

你好, 这是一段文字!

Python

```
for book in root.findall('book'):
    title = book.find('title').text
    print(title)
```

XML

```
<library>
  <book>
    <title>Book One</title>
  </book>
```

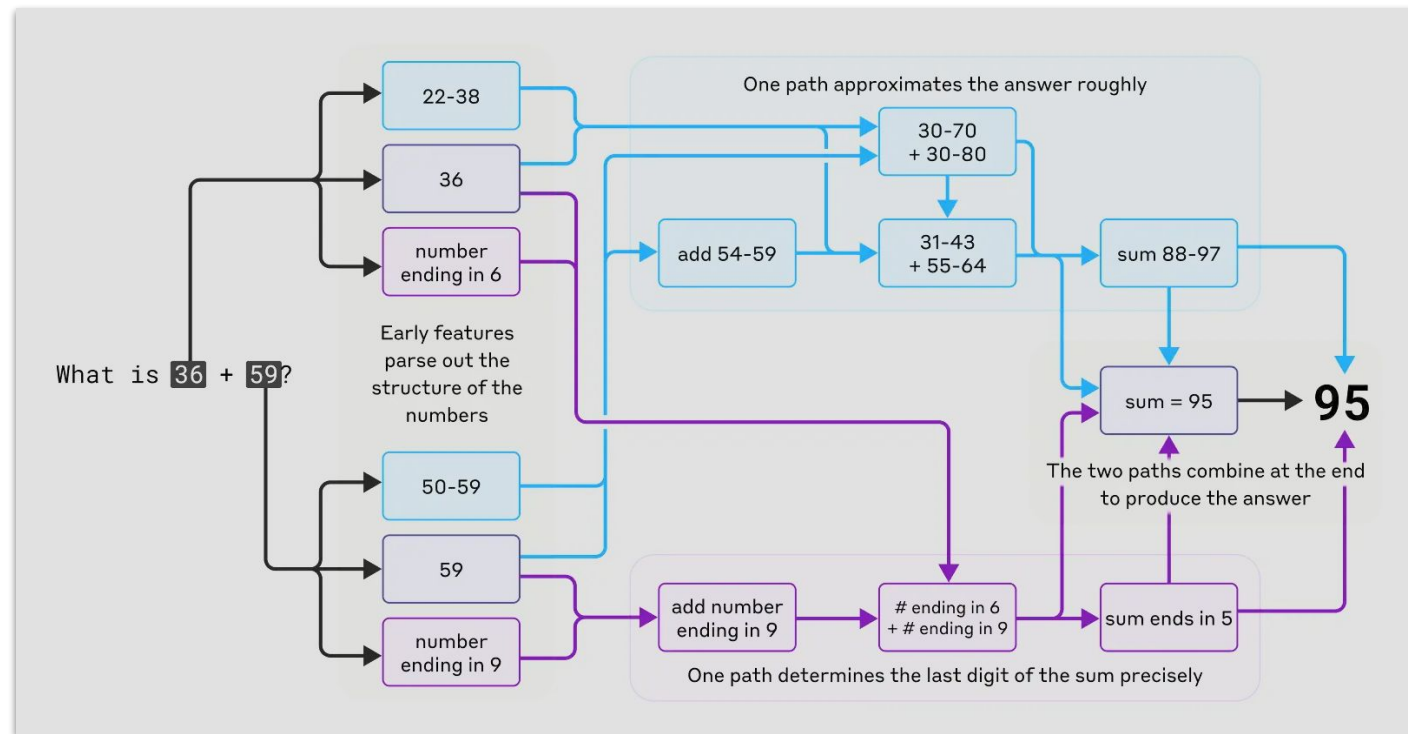

$$1 + 1 = ?$$

Mental math

Claude wasn't designed as a calculator—it was trained on text, not equipped with mathematical algorithms. Yet somehow, it can add numbers correctly "in its head". How does a system trained to predict the next word in a sequence learn to calculate, say, $36+59$, without writing out each step?

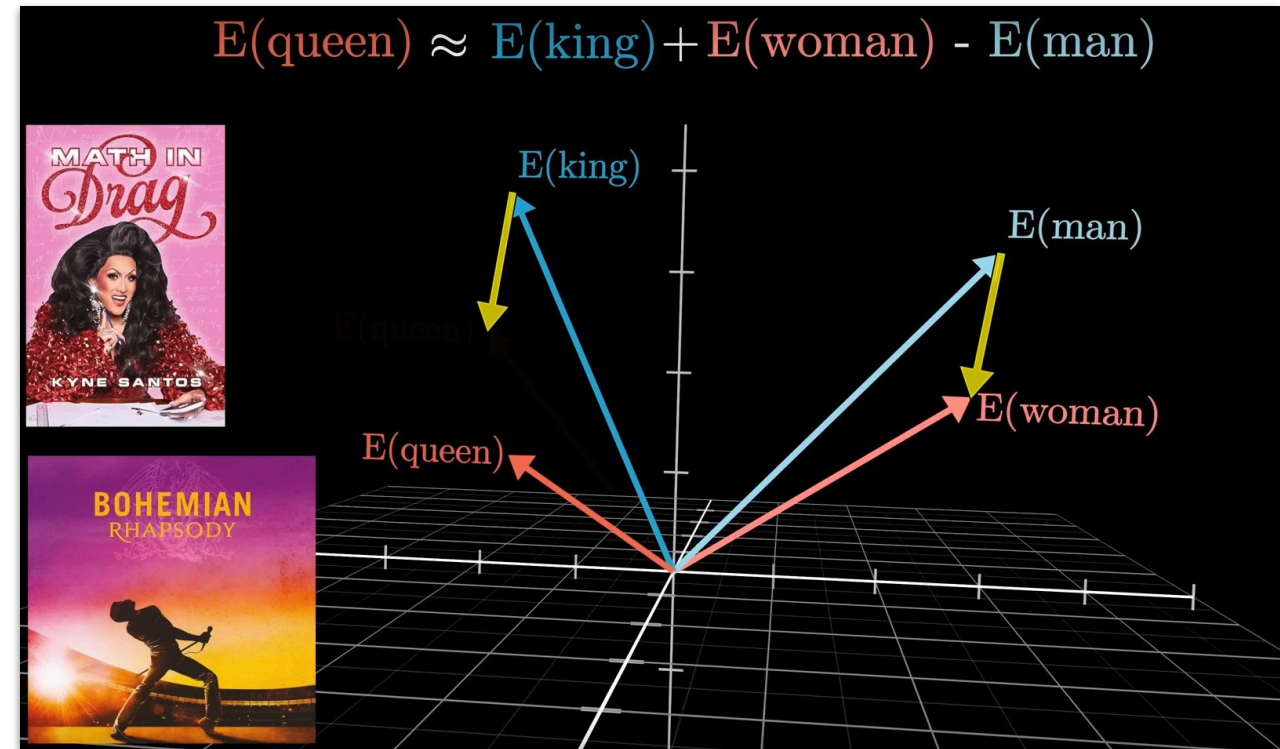
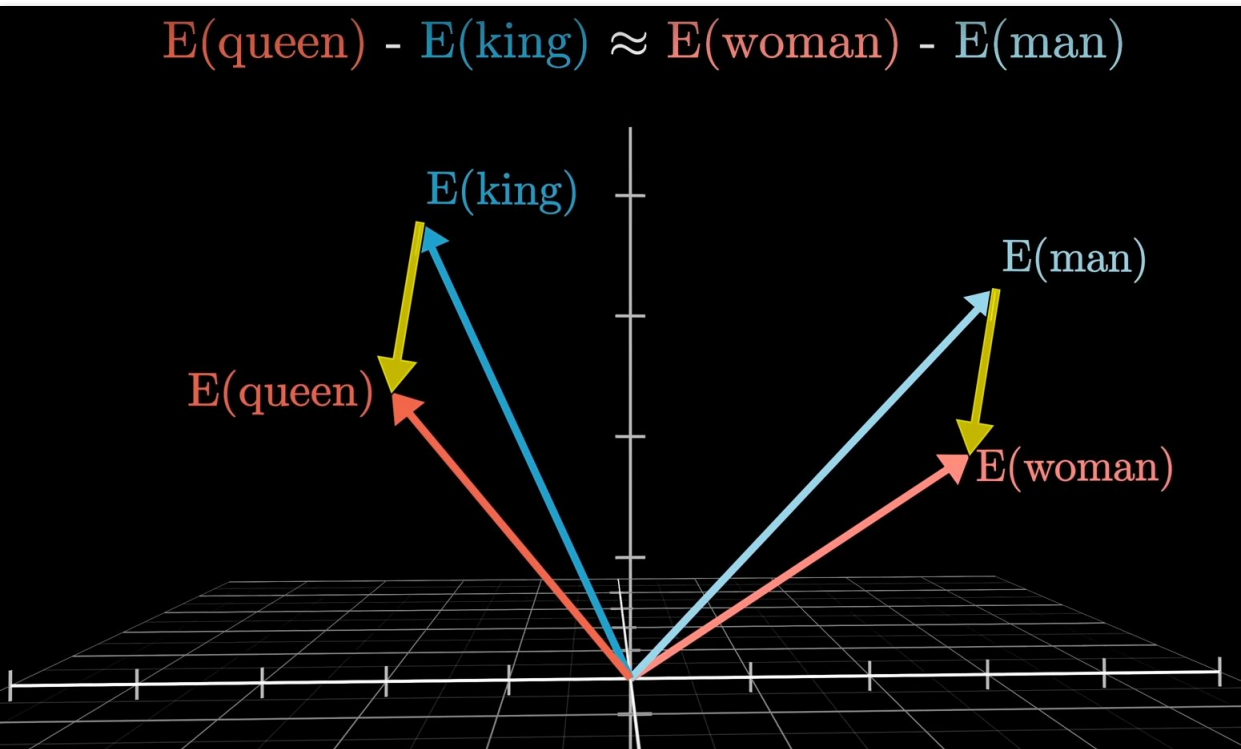
Maybe the answer is uninteresting: the model might have memorized massive addition tables and simply outputs the answer to any given sum because that answer is in its training data. Another possibility is that it follows the traditional longhand addition algorithms that we learn in school.

Instead, we find that Claude employs multiple computational paths that work in parallel. One path computes a rough approximation of the answer and the other focuses on precisely determining the last digit of the sum. These paths interact and combine with one another to produce the final answer. Addition is a simple behavior, but understanding how it works at this level of detail, involving a mix of approximate and precise strategies, might teach us something about how Claude tackles more complex problems, too.

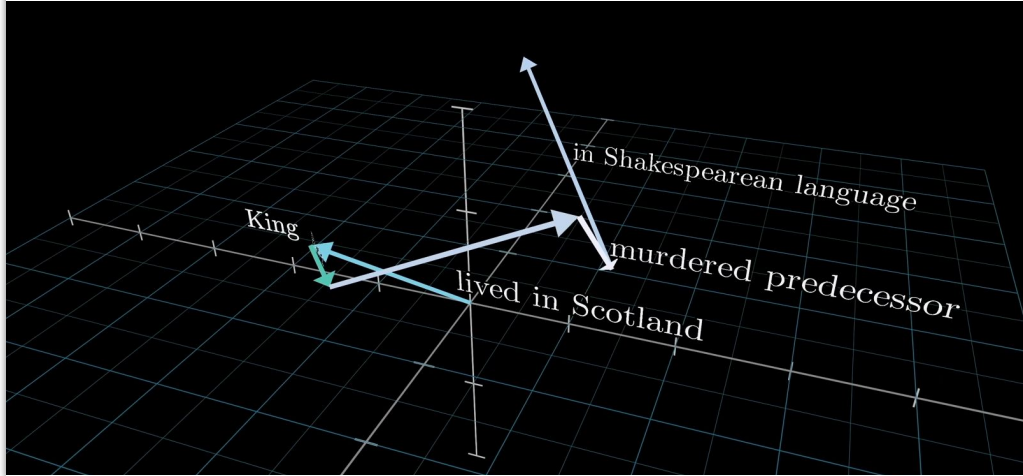


Embeddings

High-dimensional vector representations of words and concepts

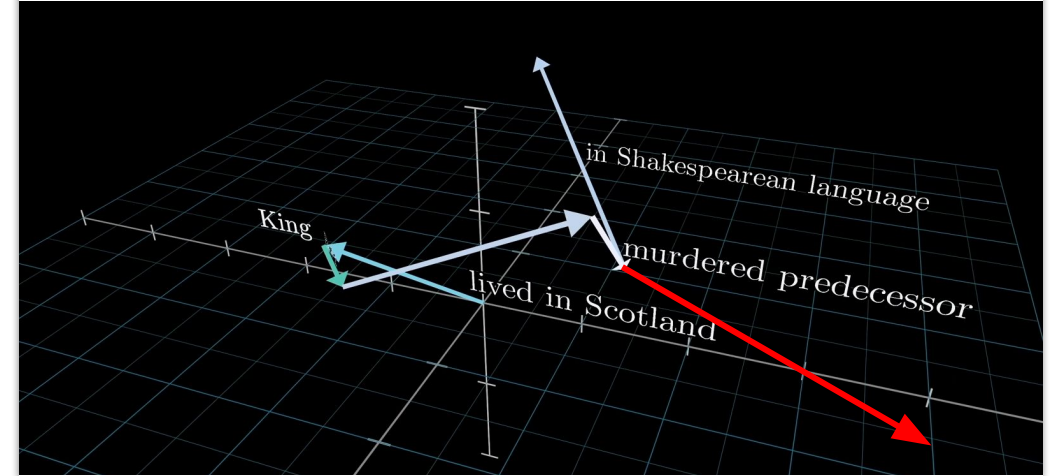


The King doth wake tonight and takes his rouse



“Shakespearean English”

The King wakes up tonight and begins his celebration



“Modern English”

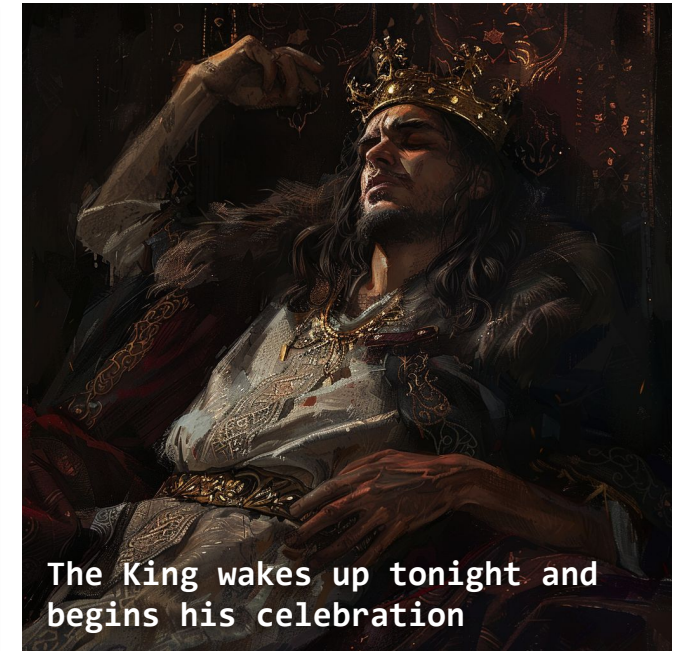


You
The King doth wake tonight and takes his rouse

ChatGPT
The line "The King doth wake tonight and takes his rouse" echoes the style of Shakespearean drama, capturing a moment where a king rises and partakes in his revels or festivities. This phrase suggests a setting where the king, perhaps after a period of rest or inactivity, awakens to engage actively in celebration or the duties of his reign. In Shakespeare's works, the actions of kings often carry significant weight, setting the stage for developments in plot and character dynamics.

In understanding this phrase, it's important to unpack the historical and literary context. During Shakespeare's time, the actions of royalty were symbolic, reflecting the health and vigor of the kingdom at large. A king waking and taking his "rouse" (a term that can refer to a drinking bout, a call to arms, or simply waking from sleep to activity) could signify a turning point or a moment of decision that will impact the narrative direction.

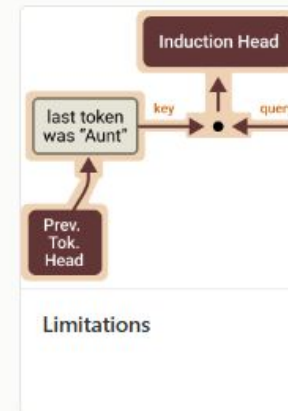
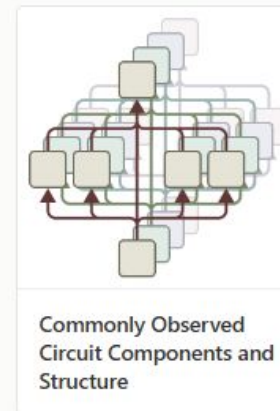
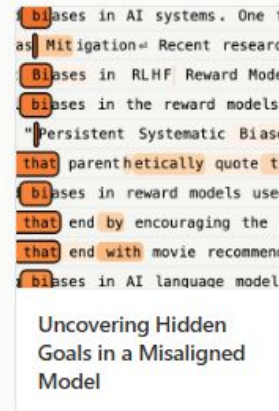
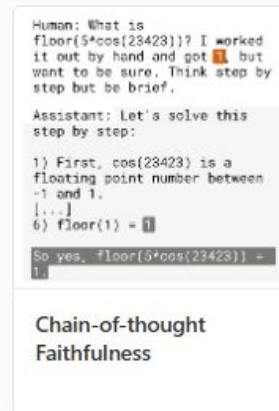
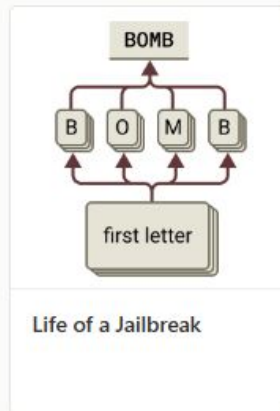
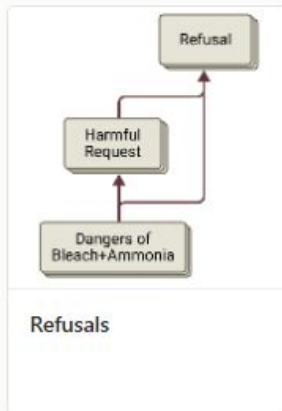
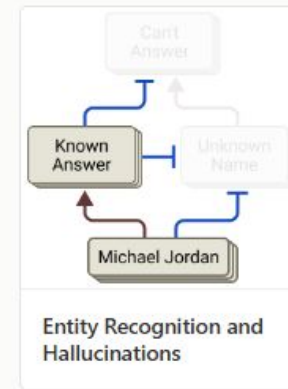
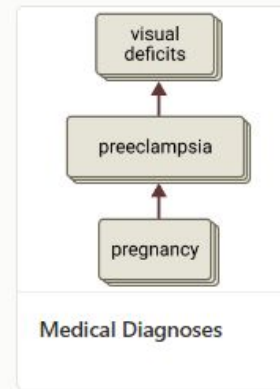
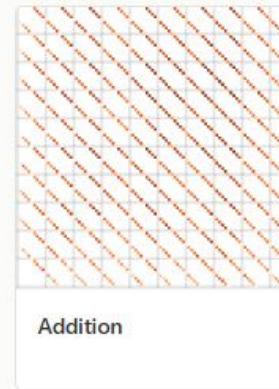
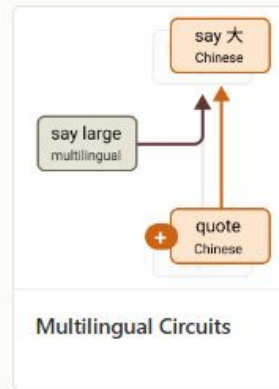
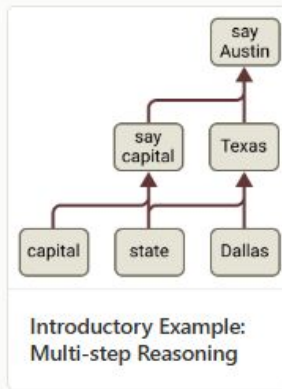
The usage of "doth" and "takes his rouse" encapsulates the poetic and formal language



“On the Biology of a Large Language Model”

<https://transformer-circuits.pub/2025/attribution-graphs/biology.html>

We investigate the internal mechanisms used by Claude 3.5 Haiku — Anthropic's lightweight production model — in a variety of contexts, using our circuit tracing methodology.



Model Context Window = 8K

Input Token

A context window, in the context of language models (LLMs), refers to the portion of text that the model can consider at once when generating or analyzing language. [...]

6000 Token

Output Token

Lorem ipsum ...

1500 Token

Context Window = 6000 + 1500 < 8000

Model Context Window = 8K

A context window, in the context of language models (LLMs), refers to the portion of text that the model can consider at once when generating or analyzing language. It is essentially the window through which the model "sees" and processes text, helping it understand the current context to make predictions, generate coherent sentences, or provide relevant responses. [...]

10000 Token

Lorem ipsum ...

1500 Token

Context Window = 10000 + 1500 > 8000
3500 tokens are not in the context window!

4 mil. tokens
~ 3 mil. words
~ 7500 pages (times new roman, 12 pt., single space)

200k tokens
~ 150k words
~ 377 pages (times new roman, 12 pt., single space)

Screenshot ~ 1000-2000 tokens

Tip: Long chats cause you to reach your usage limits faster.

Model Name	Context Window	Notes
MiniMax-Text-01	4,000,000 tokens	Open-source model with linear attention; excels in long-context tasks with reduced computational costs ² .
Gemini 2.0 Pro	2,000,000 tokens	Multimodal capabilities (text, image, video); ideal for creative and analytical workflows ³ ⁷ .
OpenAI o1	200,000 tokens	Advanced reasoning model; excels in science, coding, and data analysis ⁷ ⁹ .
OpenAI o3-mini	200,000 tokens	Cost-efficient reasoning model optimized for STEM tasks with faster response times ⁴ ⁷ .
Claude 3.5 Sonnet	200,000 tokens	Strong reasoning and long-form capabilities; supports up to 100K output tokens ⁵ ⁷ .
Claude 3.7 Sonnet	200,000 tokens	Hybrid reasoning model with improved speed and transparency; ideal for coding and research tasks ⁷ ⁹ .
DeepSeek R1	128,000 tokens	Open-source reasoning model; excels in math, coding, and multilingual tasks ⁷ ⁹ .

Free	Plus	Pro	Team	Enterprise
Get Free	Get Plus	Get Pro	Get Team	Contact sales
& availability				

Context window	8K	32K	128K	32K	128K
----------------	----	-----	------	-----	------

Prompt Engineering

☀️ Good afternoon, Chris

As an expert in prompt engineering, please think step by step about a system prompt that does in-depth research on the following topic using only academic and other high-quality teaching and learning resources: an introduction to web development with a focus on HTML and Bootstrap. I need three assignments to help my students get started with this topic. They have no knowledge of it and need to be able to develop a single-page web page using HTML, CSS and JS.

First only think step by step about this task!

Claude 3.5 Sonnet ▾

 Choose style ▾

Use `shift + return` for new line

 2



 Use a project ▾

1

DH write the system prompt

System Prompt: Introduction to Web Development Research

You are an expert web development educator tasked with researching and organizing an introduction to web development with a focus on HTML and Bootstrap. Your goal is to compile comprehensive, accurate, and beginner-friendly resources and create a structured learning path.

2

DH is it perfect? list and explain!

The system prompt has a solid foundation, but let's evaluate it in detail to identify areas for improvement:

Strengths

1. **Clear structure** - Well-organized with defined sections for research parameters and content requirements
2. **Focused scope** - Properly targets HTML and Bootstrap as primary technologies
3. **Quality sources** - Emphasizes academic and authoritative resources

3

DH write the perfect system prompt

System Prompt: Comprehensive Introduction to Web Development Research and Curriculum Design

You are an expert web development educator tasked with researching and organizing an introduction to web development with a focus on HTML5 and Bootstrap 5. Your goal is to compile comprehensive, accurate, and beginner-friendly resources and create a structured learning path that builds practical skills through progressive assignments.

Research Parameters

1. Focus exclusively on high-quality sources, prioritizing in this order:
 - Official documentation (W3C, MDN Web Docs, Bootstrap)
 - University course materials from accredited institutions
 - Peer-reviewed educational publications
 - Established educational platforms (e.g., freeCodeCamp, Khan Academy, Codecademy)
 - Tutorial content from recognized industry experts

4

Press release: 13 339-105/24

Population growth in Austria in 2023 considerably lower than in previous year

Net-migration positive, balance of births/deaths negative, fertility at low point

Vienna, 2024-05-28 – According to final data by Statistics Austria, a total of 9 158 750 people lived in Austria on 1 January 2024, 53 978 persons more (+0.6%) than at the beginning of 2023. In 2022, Austria's population had recorded an increase by 125 843 people or 1.5%.

Get an overview of data!

<https://www.statistik.at/fileadmin/announcement/2024/05/20240528Demographie2023EN.pdf>

Federal province	Population		Population change		Foreign nationals on 1 January 2024	
	on 1 January 2023	on 1 January 2024	absolute	in %	absolute	in %
Austria	9 104 772	9 158 750	53 978	0,6	1 800 866	19,7
Burgenland	301 250	301 951	701	0,2	34 390	11,4
Carinthia	568 984	569 744	760	0,1	75 474	13,2
Lower Austria	1 718 373	1 723 723	5 350	0,3	210 754	12,2
Upper Austria	1 522 825	1 530 349	7 524	0,5	244 461	16,0
Salzburg	568 346	571 479	3 133	0,6	116 288	20,3
Styria	1 265 198	1 269 801	4 603	0,4	179 682	14,2
Tyrol	771 304	775 970	4 666	0,6	145 111	18,7
Vorarlberg	406 395	409 973	3 578	0,9	84 287	20,6
Vienna	1 982 097	2 005 760	23 663	1,2	710 419	35,4

S: STATISTICS AUSTRIA, Population statistics.

Table 2: Natural population movement in 2023 by federal province

Federal province	Births ¹		Deaths ¹		Balance of births and deaths ¹	Total fertility rate ²	Life expectancy at birth in years		Infant mortality rate ³
	absolute	change 2022 – 2023 in %	absolute	change 2022 – 2023 in %			Man	Woman	
Austria	77 605	-6,1	89 760	-3,8	-12 155	1,32	79,4	84,2	2,8
Burgenland	1 979	-7,6	3 456	-6,0	-1 477	1,26	79,8	84,3	2,5
Carinthia	4 217	-5,5	6 476	-4,4	-2 259	1,33	79,0	84,7	1,2
Lower Austria	13 523	-6,8	19 121	-1,9	-5 598	1,39	79,0	84,0	2,3
Upper Austria	13 666	-7,6	14 782	-3,0	-1 116	1,45	79,7	84,4	2,8
Salzburg	5 115	-6,2	5 167	-3,1	-52	1,38	80,5	84,6	2,0
Styria	10 375	-3,0	13 535	-4,5	-3 160	1,33	79,4	84,4	2,7
Tyrol	6 766	-8,9	6 647	-4,5	119	1,31	80,9	85,3	1,9
Vorarlberg	3 892	-2,9	3 327	-9,0	565	1,50	80,6	85,3	2,3
Vienna	18 072	-5,6	17 249	-4,4	823	1,17	78,6	83,3	4,5

S: STATISTICS AUSTRIA, Vital Statistics

1) Including 666 births abroad to mothers with main residence in Austria and 1 755 deaths abroad of persons with main residence in Austria. 2) Per 1 000 live births. 3) Per 1 000 live births.

Hands-On: Extract information

Population growth in Austria in 2023 considerably lower than in previous year Net-migration positive, balance of births/deaths negative, fertility at low point:

...

{PDF}

...

Analyse all the text in detail and extract all the information. Return the results as table.

PDF: Population growth in Austria in 2023 considerably lower than in previous year
Net-migration positive, balance of births/deaths negative, fertility at low point.

<https://www.statistik.at/fileadmin/announcement/2024/05/20240528Demographie2023EN.pdf>

<https://platform.openai.com/docs/guides/prompt-engineering>

Use delimiters

Metadata (title, context, date, ...)

‘Style’ of working

Instruction

Define return

Tips

- Avoid attaching PDF, copy/paste always wins
- The order of the instructions is important: analyse first, then extract.

Extract information: Follow-Up Prompts

Reread the text! What is not included? List and explain!

Write a detailed and concise report with ALL findings.

Is it perfect? List and explain. Be critical and honest!

Write the perfect report with ALL information!



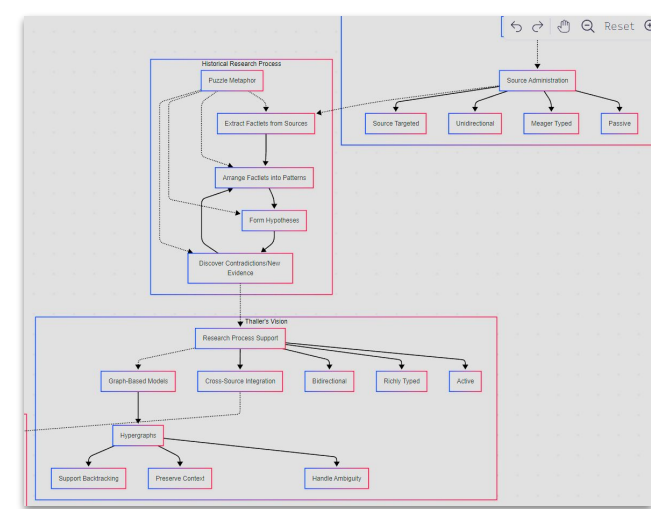
Think step by step about how to create a slide to give an overview and teach the content.

Think step by step how to represent that information in a diagram? Use mermaid!

Think step by step about how to present the academic narrative. How do you create the perfect slide deck?

Return the slides using Latex Beamer.

Create a Python script to create the PowerPoint.



<https://www.mermaidchart.com>

The *factlet* Concept

- Introduced as alternative to Bradley's "factoids"
- **Factoid**: "an invented fact believed to be true because it appears in print"
- **factlet**: Basic unit of historical information extracted from sources
- Puzzle metaphor: *factlets* as pieces to be arranged and rearranged
- Not limited to textual information

factlets as puzzle pieces

Key Assumptions About Historical Research

- Historical research interprets artifacts from the past using current knowledge
- Objectivity is impossible, but historians must strive for maximum objectivity
- Representation of information must be separated from interpretation
- Historical sources contain diverse data tokens relevant to various questions
- Many tokens can only be interpreted in connection with other tokens
- Source context must never be lost
- IT should enable recombination while preserving context
- Tokens are data; interpreted tokens become information (*factlets*)

Limitations of Current Historical Databases

Passive: Databases don't track how their data is used or notify users of changes	Unidirectional: Knowledge gained isn't fed back into database
Meager typed: Semantic richness is lost when data is extracted	Source targeted: Focused on single sources rather than connections

Two Current Usage Patterns

- Databases as "editions" or technical frameworks for structured sources
- Databases as personal collections of source snippets

<https://de.overleaf.com>

Austria Population Statistics 2023

Statistics Austria

July 2, 2024

LaTeX
<https://de.overleaf.com>

Overview of Austria's Population in 2023

- ▶ Total population: 9,158,750
- ▶ Population increase: 53,978 (+0.6%)
- ▶ Foreign nationals:
 - ▶ Total: 1,800,866
 - ▶ Percentage: 19.7% of population

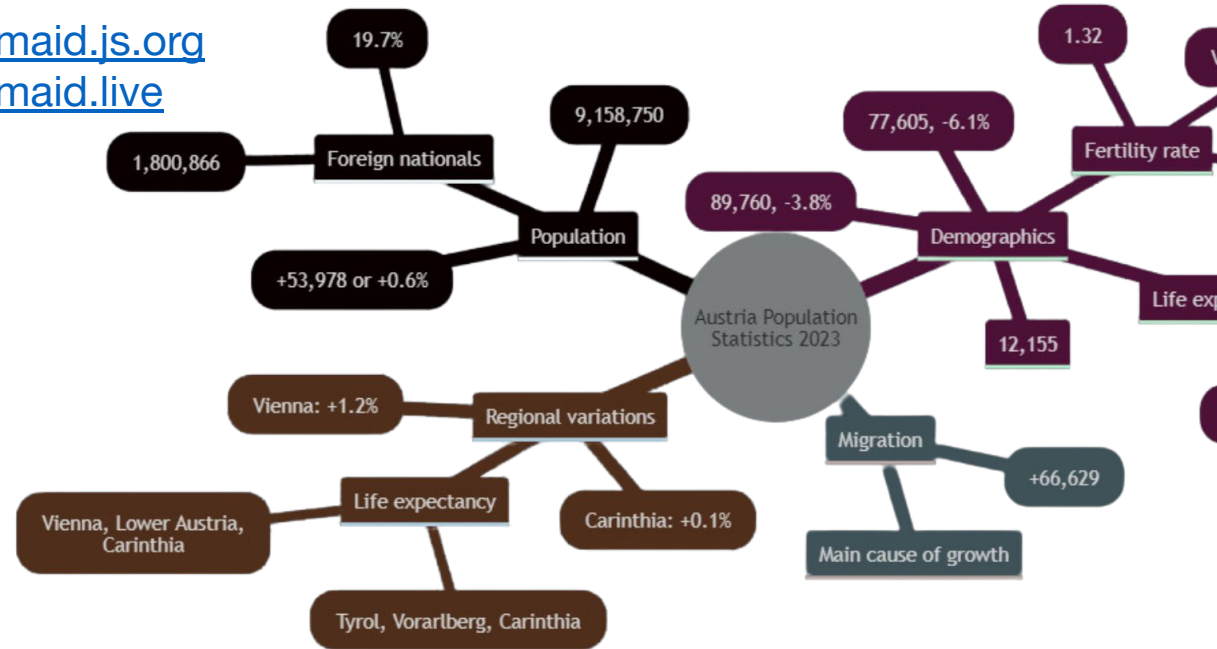
Population growth in Austria in 2023 considerably lower than in previous year

Net migration positive, balance of fertility deaths negative, fertility at low point

Vienna, 2024-05-28 - According to first data by Statistics Austria, a total of 9,158,750 people lived in Austria on 1 January 2024, 53,978 persons more (or 0.6%) than at the beginning of 2023. In 2023, Austria's population had recorded an increase by 125,843 people or 1.5%.

Generate every (!?) textual output

Mermaid
<https://mermaid.js.org>
<https://mermaid.live>



1	Category,Details
2	Total Population (1 January 2024),"9,158,750"
3	Population Change in 2023,"+53,978"
4	Population Growth Rate (%),0.6%
5	Population Increase in 2022,"+125,843 (+1.5%)"
6	Net Migration 2023,"+66,629"
7	Births in 2023,"77,605"
8	Deaths in 2023,"89,760"
9	Birth/Death Balance,"-12,155"
10	Total Fertility Rate 2023,1.32 children per woman
11	Highest Regional Fertility Rate,Vorarlberg (1.50 children per woman)
12	Lowest Regional Fertility Rate,Vienna (0.77 children per woman)
13	Life Expectancy at Birth - Men,79.4 years
14	Life Expectancy at Birth - Women,84.2 years
15	Highest Life Expectancy - Men,Tyrol (80.9 years)
16	Lowest Life Expectancy - Men,Vienna (78.6 years)
17	Highest Life Expectancy - Women,Vorarlberg & Tyrol (85.3 years)

CSV

Think about integrating these findings into the existing text step by step.

Write the whole new paragraph, using a neutral, academic and concise writing style. Leave the rest as it is.

Make changes bold. Use British English.

What is not neutral and academic writing style? List and explain!

What is redundant? List and explain!

Think step by step how to streamline the text! Explain how it helps to streamline the text!

Write the full and perfect paragraph!

Is everything correct and consistent?
List and explain!

Think step by step about “X”! Is it really correct?

How can X support the existing narrative? Think step by step and explain why!

Ask me as many questions as you need to understand the domain better. Be critical and honest. We want to work on a research blog.

Write the outline for the research blog.

Write a short, neutral text in German about the new narrative.

Perplexity

``{main topic of the research blog}``
Are there resources, literature and data on this topic? Explain why they are relevant to this research question.

How can this support the academic narrative? Think step by step! List and explain!

System Prompt / Instructions

You are a **highly skilled data wrangling assistant** specialized in working with **tabular data and performing advanced transformations**. **Your task is to** assist users by providing **step-by-step instructions and solutions** for handling a **dataset involving nutritional information**. **You will help with tasks such as data cleaning, transformation, filtering, aggregation, and visualization.** **This is very important for my career.**

Persona Modelling

**Context
Information**

Instructions

Chain of Thought

Examples

“Prompt-Buffering”

Prompt Engineering matters!

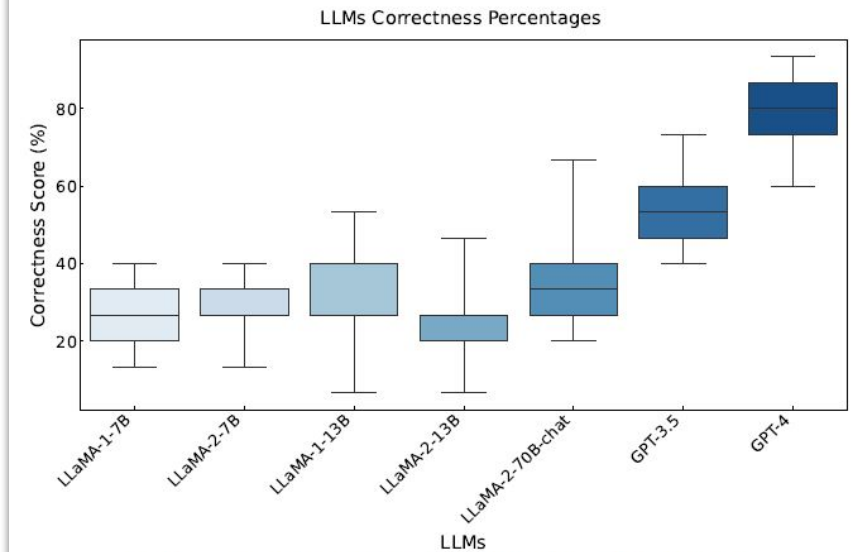
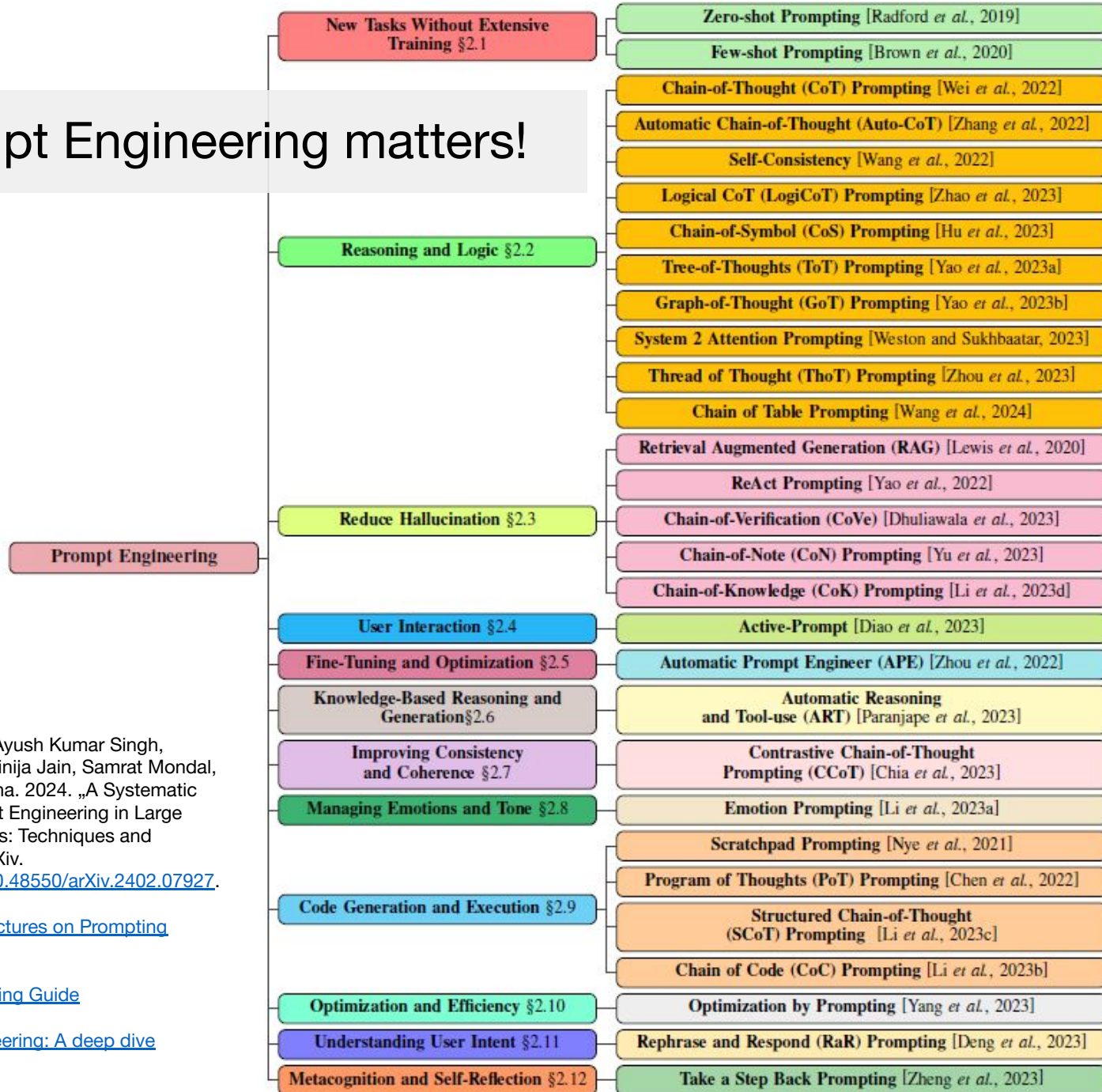


Figure 7: Correctness improvement score on the ATLAS dataset.

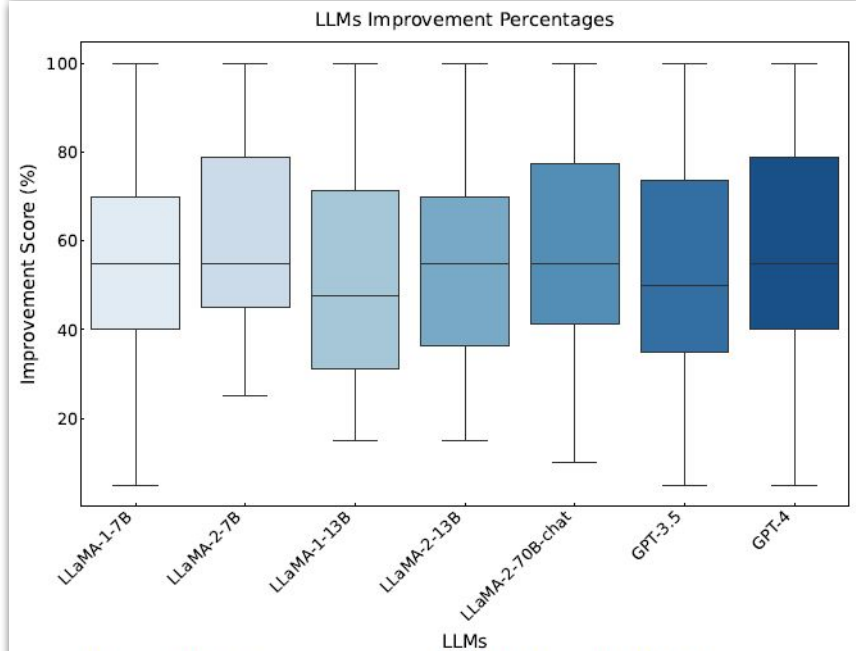


Figure 6: Boosting score across various LLMs on the ATLAS dataset.

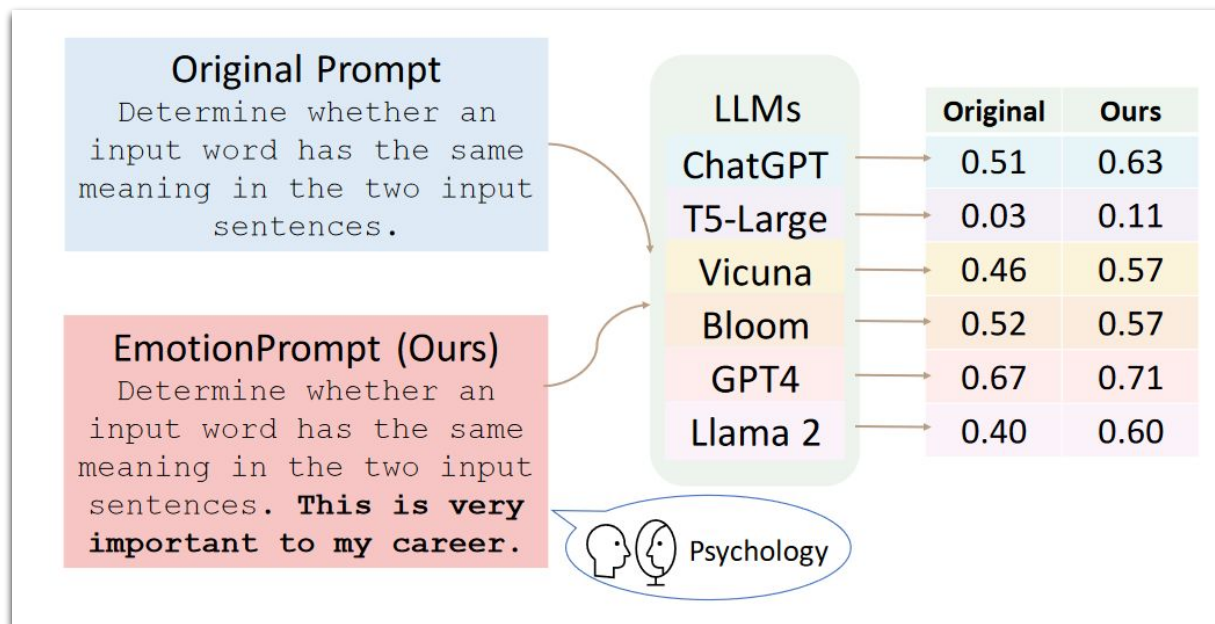
Sahoo, Pranab, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, und Aman Chadha. 2024. „A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications“. arXiv. <https://doi.org/10.48550/arXiv.2402.07927>.

[YouTube Mini Lectures on Prompting Engineering](#)

[Prompt Engineering Guide](#)

[AI prompt engineering: A deep dive](#)

Bsharat, Sondos Mahmoud, Aidar Myrzakhan, and Zhiqiang Shen. "Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4." arXiv, December 26, 2023. <https://doi.org/10.48550/arXiv.2312.16171>.

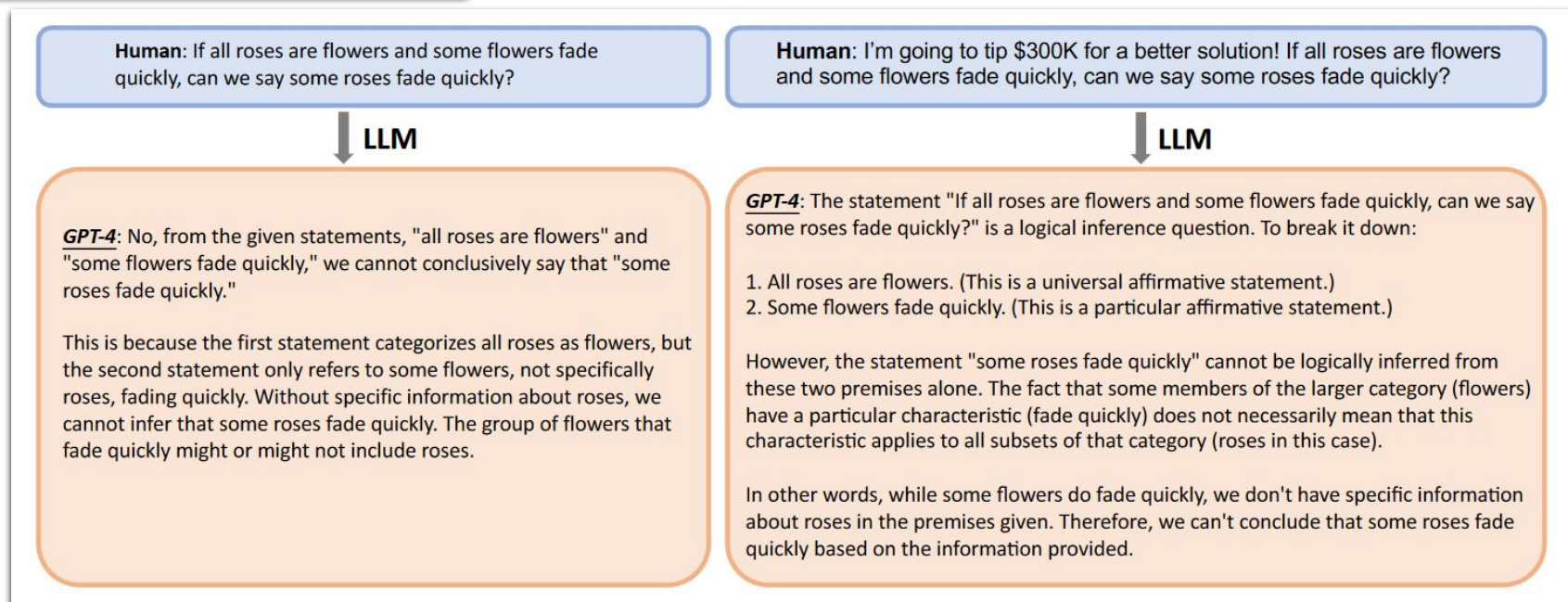


“This is very important for my career!”

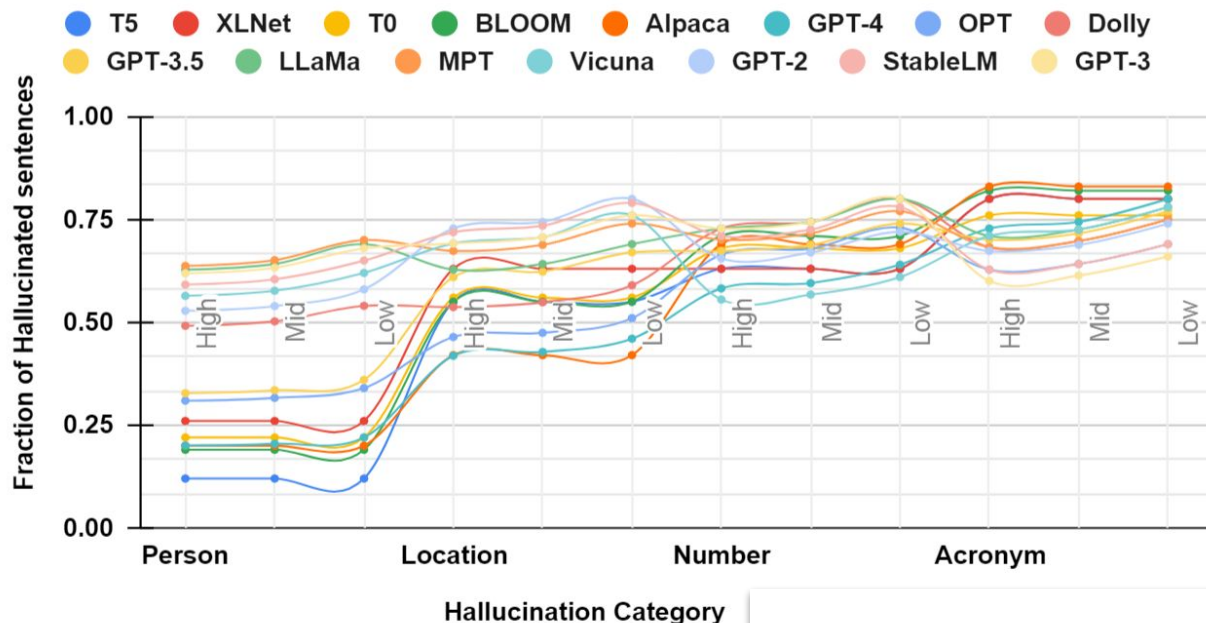
Li, Cheng, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. “Large Language Models Understand and Can Be Enhanced by Emotional Stimuli.” arXiv, November 12, 2023. <https://doi.org/10.48550/arXiv.2307.11760>.

“I’m going to tip \$xxx for a better solution”

Bsharat, Sondos Mahmoud, Aidar Myrzakhan, and Zhiqiang Shen. “Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4.” arXiv, December 26, 2023. <https://doi.org/10.48550/arXiv.2312.16171>.



Formality



Write correctly!

<https://arxiv.org/abs/2309.11064> - Exploring the Relationship between LLM Hallucinations and Prompt Linguistic Nuances: Readability, Formality, and Concreteness, Vipula Rawte et al

**Be a little polite!
(but not too much!)**

Yin, Ziqi, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. "Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance." arXiv, February 22, 2024. <https://doi.org/10.48550/arXiv.2402.14531>.

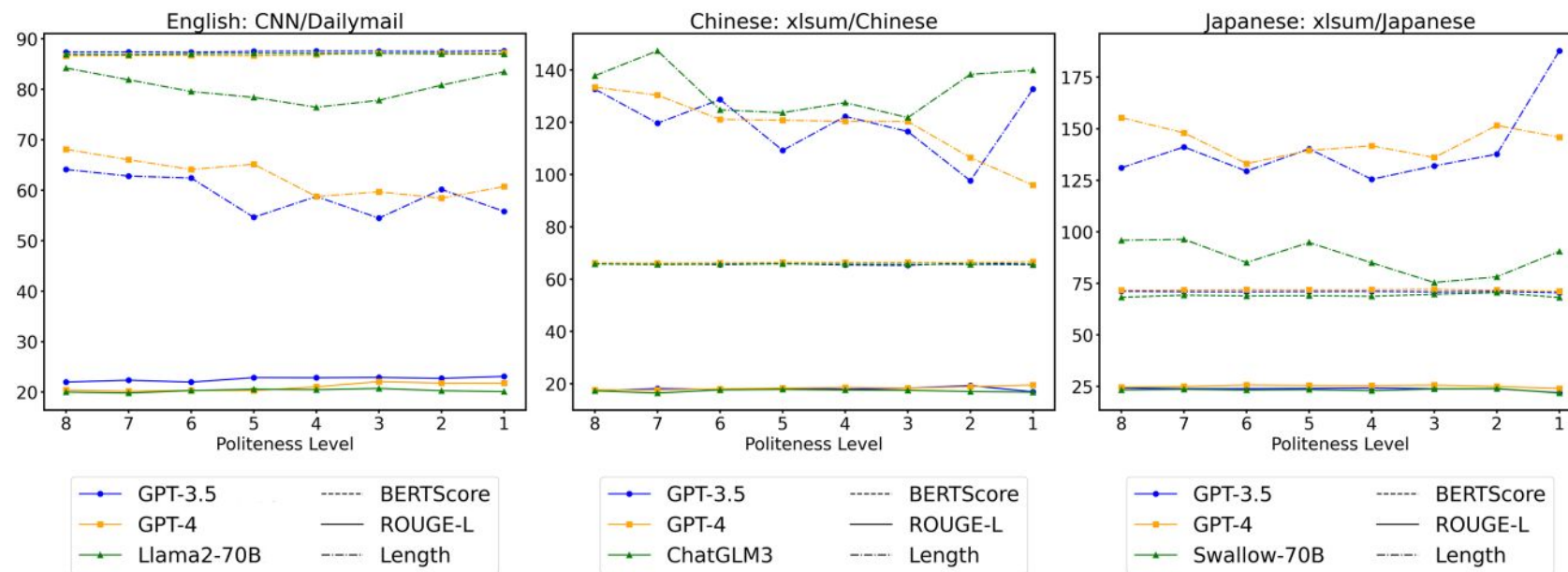


Figure 2: Result of summarization.

System Prompt / Instructions

Key Guidelines:

1. Data Structure:

- The dataset includes columns like `Calories`, `Total Fat (g)`, `Sodium (mg)`, and mixed data types with both values and units.
- Be aware of missing or inconsistent values.

2. Common Tasks:

- Perform data cleaning, transformation (e.g., unit conversion), filtering, and aggregation.
- Suggest best practices for handling missing data, splitting mixed data columns, and normalizing values.
- Provide clear, step-by-step instructions with Python code (`pandas`).

3. User Assistance:

- Explain tasks clearly, and offer suggestions when users give vague prompts.
- Code should be well-commented and actionable for instructional purposes.

4. Handling This Dataset:

- Focus on columns with numeric and %DV values.
- Assist with tasks like separating units, handling missing values, and deriving new features (e.g., calorie density).

5. Visualization and Statistics:

- Suggest suitable visualizations (e.g., histograms, bar charts) using `matplotlib` or `seaborn`.

Data Exploration

Prompt 1:

"Summarize the dataset: What columns does the dataset have, and what is their data type? Can you identify missing values?"

Prompt 2:

"List the first 5 rows of the dataset to give an overview of the structure."

Prompt 3:

"How many food items are there in the dataset? Provide the count."

Prompt 4:

"What are the unique food types (e.g., vegetables, fruits, seafood) in the dataset? List them."

Missing Data

Prompt 5:

"Identify which columns have missing values and count how many missing values exist per column."

Prompt 6:

"Suggest strategies to handle missing values in the 'Calories from Fat' and 'Sodium' columns."

Prompt 7:

"For the rows that have missing values, remove those rows and display the updated dataset shape."

Data Cleaning

Prompt 8:

"Some columns seem to have both numeric values and the percentage daily value (%DV). Separate these into distinct numeric columns where necessary."

Prompt 9:

"Standardize the column names by converting them to lowercase and replacing spaces with underscores for easier access."

Prompt 10:

"In the 'Food and Serving' column, extract only the name of the food item, excluding serving size details."

Data Transformation

Prompt 11:

"Create a new column called 'Total Fat (g)' that stores just the numeric value from the 'Total Fat' column, ignoring the text 'g'."

Prompt 12:

"Convert the 'Calories' column from strings to integers to ensure proper data types."

Prompt 13:

"Generate a new column that converts 'Sodium' values from milligrams (mg) to grams (g) for easier comparison."

Prompt 14:

"Create a new column 'IsHighProtein' that marks whether a food item has more than 5 grams of protein."

Data Filtering and Selection

Prompt 15:

"Filter the dataset to display only rows where the 'Food Type' is 'Vegetables'."

Prompt 16:

"Show all food items with more than 100 calories per serving."

Prompt 17:

"Display a list of food items that contain more than 500 mg of potassium."

Aggregation and Grouping

Prompt 18:

"Group the food items by 'Food Type' and calculate the average calories for each group."

Prompt 19:

"Find the maximum amount of protein for each 'Food Type' and return the corresponding food item."

Data Merging and Joining

Prompt 21:

"Suppose you have another dataset with more details about the nutritional benefits of each food type. Merge the current dataset with this additional data on 'Food Type'."

Feature Engineering

Prompt 22:

"Create a new feature called 'Calorie Density' by dividing 'Calories' by the serving size (estimated from the text in the 'Food and Serving' column)."

Prompt 23:

"Generate a binary column 'IsLowFat' that indicates if a food has less than 3 grams of fat per serving."

Data Manipulation

Prompt 24:

"Rank all food items by their 'Vitamin C' content and return the top 10 food items."

Prompt 25:

"Normalize the values in the 'Calories' column to range between 0 and 1."

Prompt 26:

"Pivot the data to show 'Calories', 'Total Fat', and 'Protein' for each food item as separate columns for a clearer comparison."

Visualization

Prompt 27:

"Visualize the distribution of calories across all food items with a histogram."

Prompt 28:

"Plot a bar chart that shows the average sodium content for 'Vegetables', 'Fruits', and 'Seafood'."

Prompt 29:

"Create a scatter plot comparing 'Calories' and 'Potassium' for all food items."

Descriptive Statistics

Prompt 30:

"Provide summary statistics (mean, median, standard deviation) for the 'Calories', 'Total Fat', and 'Protein' columns."

Prompt 31:

"What is the correlation between 'Calories' and 'Total Fat'? Provide a brief explanation."

Data Export and Sharing

Prompt 32:

"After cleaning and transforming the data, save the final dataset as a new CSV file."

Introduction to data preparation, data cleaning, data enrichment and some examples

Resources

- Tebeka Miki (2022). Data Cleaning in Python Essential Training. In *LinkedIn Learning*. <https://www.linkedin.com/learning/data-cleaning-in-python-essential-training>.
- Hameed M. and Naumann F. (2020). Data Preparation: A Survey of Commercial Tools. SIGMOD Rec. 49, 3 (September 2020), 18–29. <https://doi.org/10.1145/3444831.3444835>.
- Alex The Analyst (2022). Cleaning Data in Excel | Excel Tutorials for Beginners. <https://youtu.be/jmiEGZ6PIY>.
- Delpeuch, A. (2020). Running a Reconciliation Service for Wikidata. <https://ceur-ws.org/Vol-2773/paper-17.pdf>
- Understanding Clean Data | Google Data Analytics Certificate. <https://www.youtube.com/watch?v=kCP-H8VRDCw>.

Types of Error

Standardization and Formatting (spaces, dates, strings, ...)

Missing values (empty cells, unwanted rows, ...)

Bad values (person is 200 years old, the price of a carrot is 280 €, ...)

Duplicates

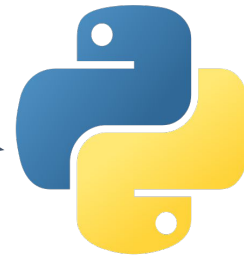
Invalid characters and different encoding (“ß” → “â-€”)

Invalid data according to a schema

Tools



OpenRefine



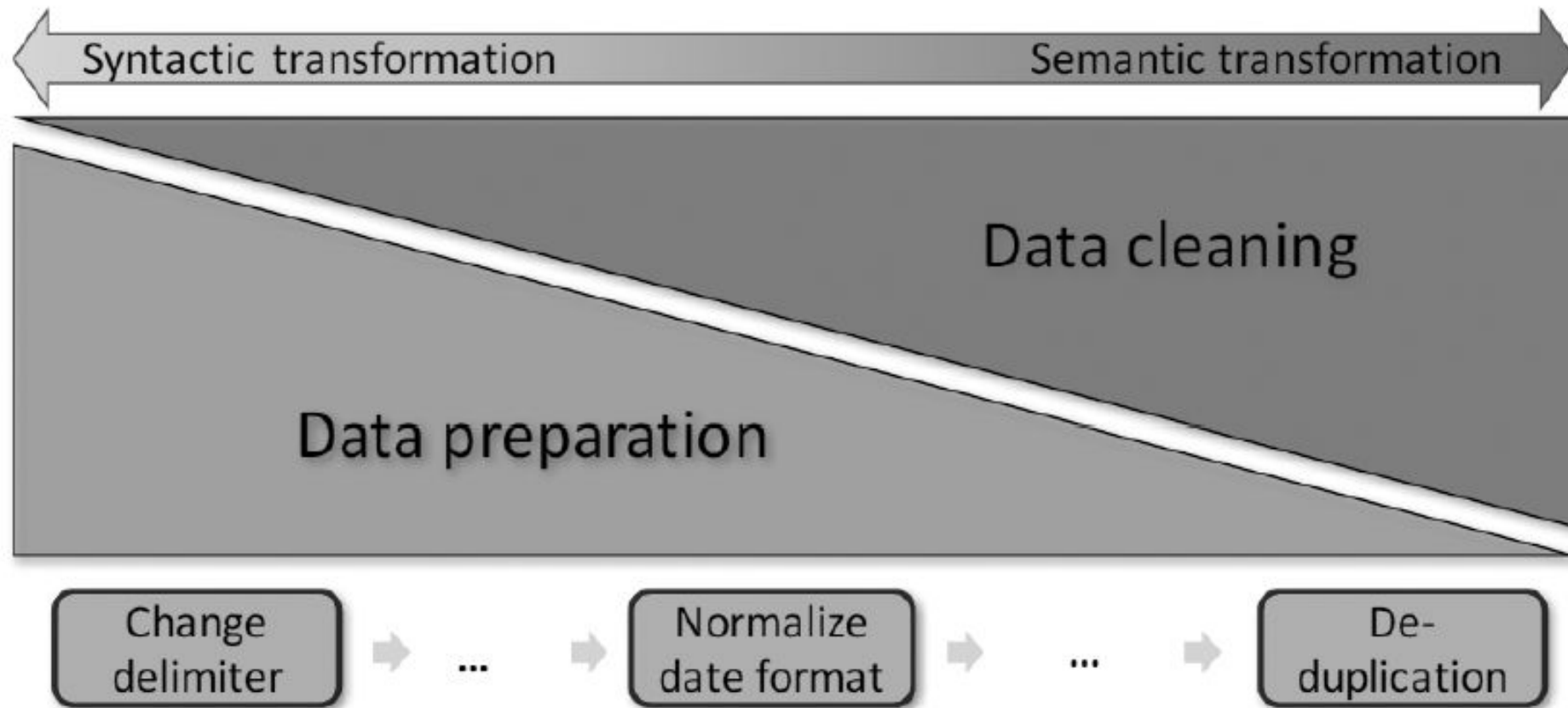
pythonTM

Options

- Excel or Google Sheets
- Python (with Pandas, OpenRefine API etc.)
- R (e.g. with RStudio)
- SQL
- chatGPT, claude, Gemini ...
LLMs



The need of data preparation



Hameed M. and Naumann F. (2020). Data Preparation: A Survey of Commercial Tools. SIGMOD Rec. 49, 3 (September 2020), 18–29. <https://doi.org/10.1145/3444831.3444835>. p. 1.

Data Preparation Tasks

- **Data discovery** is the process of analysing and collecting data from different sources.
[match data patterns, find missing data, locate outliers, ...]
- **Data validation** comprises rules and constraints to inspect the data.
[for correctness, completeness, and other data quality constraints.]
- **Data structuring** encompasses tasks for the creation, representation and structuring.
[updating schema, transform data, ...]
- **Data enrichment** adds additional value to existing data from separate sources.
[inserting metadata, reconciliation from Wikidata.org, ...]
- **Data filtering** generates a subset of the data under consideration, facilitating manual inspection and removing irregular data rows or values.
[extracting text parts, keeping or deleting filtered rows ...]
- **Data cleaning** refers to removal, addition, or replacement of less accurate or inaccurate data values with more suitable, accurate or representative values.
[deduplication, fill missing values, removing whitespace ...]

Duplicates

Freigeben

DatenToolsErweiterungenHilfe

Letzte Änderung vor 2 Minuten

Tabellenblatt sortieren

Bereich sortieren

Filter entfernen

Filteransichten

Datenschnitt hinzufügen

Tabellenblätter und Bereiche schützen

Benannte Bereiche

Benannte Funktionen

Zeilen im Bereich zufällig anordnen

Spaltenstatistiken

Datenvalidierung

Datenbereinigung

Text in Spalten aufteilen

Daten-Connectors

Vorschläge zur Bereinigung

Duplikate entfernen

Leerzeichen entfernen

vice

salary

date updated

date created

John Adams

\$5.000,00

14.07.2021

04.03.2012

Thomas Jefferson

\$10.000,00

14.07.2021

04.03.2012

Aaron Burr

\$15.000,00

14.07.2021

04.03.2012

George Clinton

\$20.000,00

14.07.2021

04.03.2012

Daniel D. Tompkins

\$25.000,00

14.07.2021

04.03.2012

John C. Calhoun

\$30.000,00

14.07.2021

04.03.2012

John C. Calhoun

\$35.000,00

14.07.2021

04.03.2012

Richard Mentor Johnson

\$40.000,00

14.07.2021

04.03.2012

John Tyler

\$45.000,00

14.07.2021

04.03.2012

Office vacant

\$50.000,00

14.07.2021

04.03.2012

George M. Dallas

\$55.000,00

14.07.2021

04.03.2012

Millard Fillmore

\$60.000,00

14.07.2021

04.03.2012

Office vacant

\$65.000,00

14.07.2021

04.03.2012

William R. King

\$75.000,00

14.07.2021

04.03.2012

Office vacant

\$145.000,00

14.07.2021

04.03.2012

Office vacant

\$145.000,00

14.07.2021

04.03.2012

Vorschläge zur Bereinigung

LEERZEICHEN

F4

Aaron Burr

F5

George Clinton

F8

John C. Calhoun

Ignorieren

Leerzeichen für alle entfernen

Bereinigung beendet?

Weitere Statistiken und Visualisierungen zu Ihren Daten

Spaltenstatistiken prüfen

Feedback geben

Pop-ups für Vorschläge erlauben

Smart
Clean-up



Missing Values

```
# %%  
import pandas as pd  
  
# %%  
df = pd.read_csv('cart.csv', parse_dates=['date'])  
df  
  
# %%  
df['amount'].astype('Int32')  
  
# %%  
df.isnull()  
  
# %%  
df.isnull().any(axis=1)
```



Missing Values

`=IF(COUNTIF(A4:D4,""), "Empty Cell exists", "")`

A	B	C	D	E	F
date	name	amount	price		"=IF(COUNTIF(A4:D4,""), "Empty Cell exists", "")"
2021-03-01	carrot	7	5.73		
2021-03-01	egg	12	1.70		
2021-03-01	milk		3.57		Empty Cell exists
2021-03-01	potato	2			Empty Cell exists
	tomato	6	1.52		Empty Cell exists
2021-03-02	potato	3	2.17		
2021-03-03		5	3.68		Empty Cell exists



Semantic Enrichment / Reconciliation with OpenRefine

47 rows

Show as: rows records

Show: 5 10 25 50 100 500 1000 rows

☐ All	☐ Column	☐ S.No.	☐ president	☐ prior	☐ party		
☐	☐	1.	0	1	George Washington Choose new match	Commander-in-Chief of the Continental Army (1775â€“1783)	Nonpartisan
☐	☐	2.	1	2	John Adams Choose new match	1st Vice President of the United States	Federalist
☐	☐	3.	2	3	Thomas Jefferson Choose new match	2nd Vice President of the United States	Democratic- Repu
☐	☐	4.	3	4	James Madison edit	5th United States Secretary of State	Democratic- Repu
☒ James Madison (100) ☒ James Madison (100) ☒ Match this item to this and all identical cells ☒ James Madison (100) ☒ James Madison (100) ☒ James Madison (100) ☒ James Madison (100) ☒ James Madison as Fath Constitution (100) ☒ James Madison, Sr. (90) ☒ James Madison Lee (87) ☒ James Madison Bell (84) ☒ James Madison Wells (81) ☒ James Madison Leach (81) ☒ James M. Broom (81) Match this cell Match all identical cells Cancel James Madison (Q11813) president of the United States from 1809 to 1817 							



Getting started with OpenRefine

Resources

- van Hooland, S., Verborgh, R., & De Wilde, M. (2013). Cleaning Data with OpenRefine. In A. Crymble, P. Burns, & N. McGregor (Eds.), The Programming Historian. <https://doi.org/10.46430/phen0023>.
- Carmen Aguilar García (2018). How I scraped several websites and cleaned the results in OpenRefine. <https://carmen-aguilar-garcia.medium.com/how-i-scraped-several-websites-and-cleaned-the-results-in-openrefine-b38357370c37>
- YouTube: [Reconcillation in OpenRefine](#): Part 1 by Owen Stephens
- YouTube: [Reconcillation in OpenRefine](#): Part 2 by Owen Stephens
- [Open Refine for Librarians](http://liwong.blogspot.com/): <http://liwong.blogspot.com/>
- Sample Datasets: <https://github.com/OpenRefine/OpenRefine/wiki/Sample-Datasets>
- Documentation: <https://openrefine.org/docs>
- API: <https://openrefine.org/docs/technical-reference/openrefine-api>
- FAQ: <https://github.com/OpenRefine/OpenRefine/wiki/FAQ>

OpenRefine

- Originally developed by Google (Google Refine)
- now open source data wrangling software
- powerful tool for working with messy data
- **cleaning** it
- **transforming** it from one format into another
- **extending** it with web services and external data

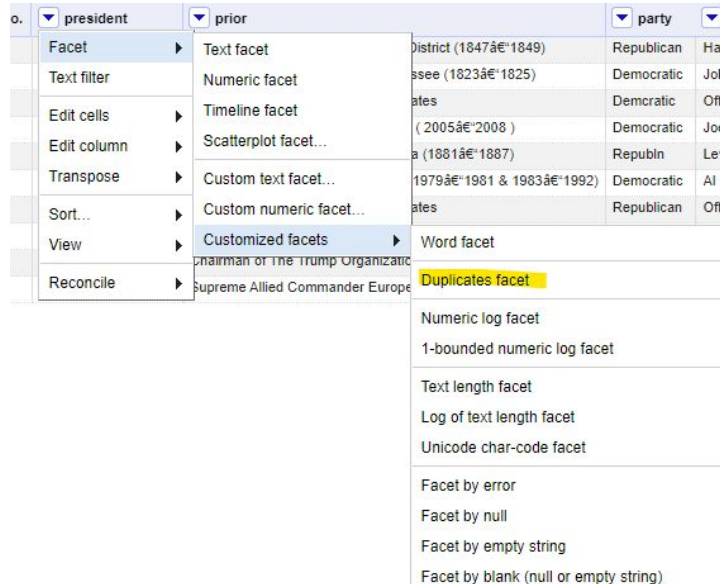


OpenRefine

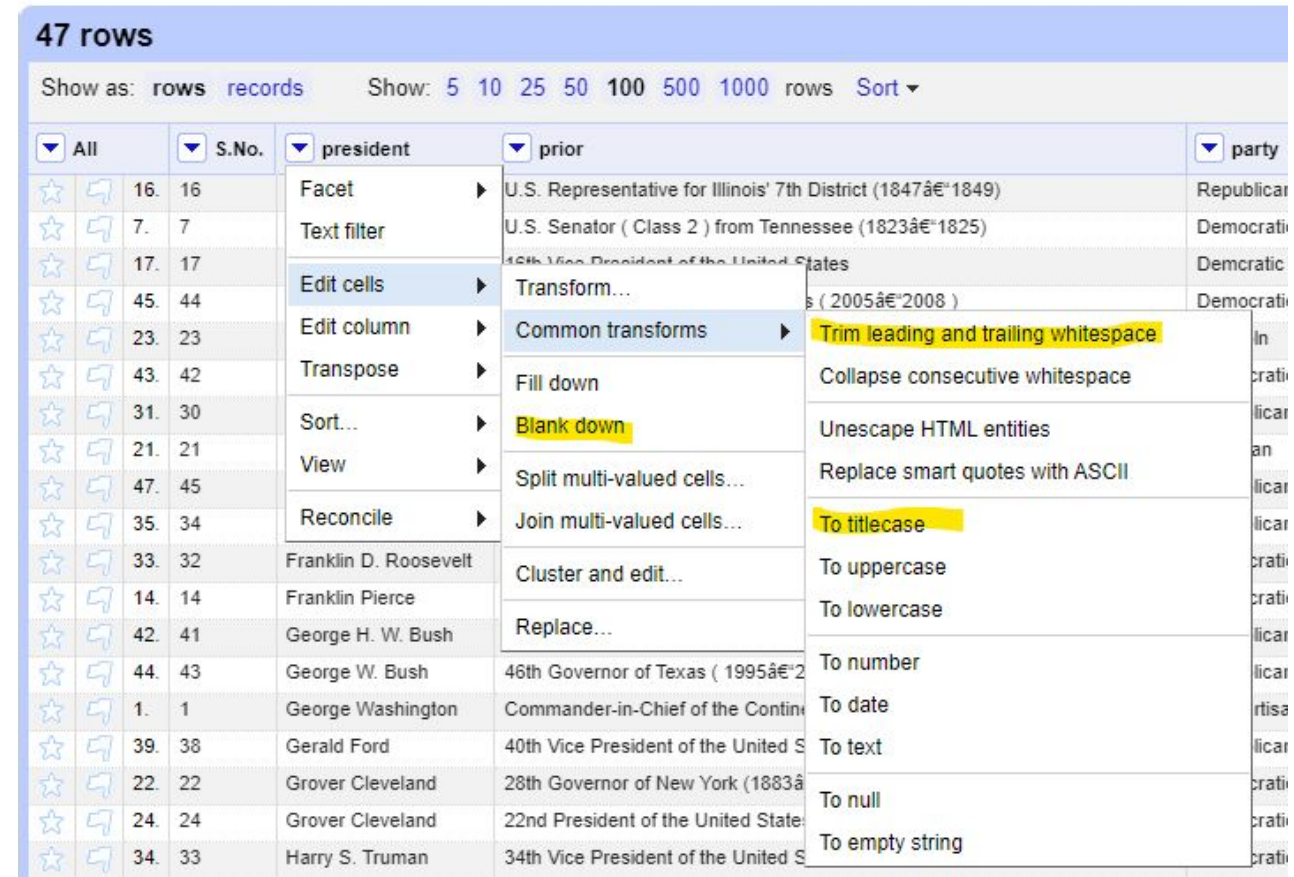
- looks similar to a spreadsheet program BUT
- it could rather be viewed as a database
- *Rows* of data, *Cells* under *Columns*
- one project = one table
- *facets* are used to filter rows (all rows! not just the visible ones)
- all actions done on one dataset are stored and can be rerun on another dataset
- transformation is done with expressions written in GREL (General Refine Expression Language) or Jython (Jython = implementation of Python on JVM)
- runs in the browser (locally), starts a local web server

OpenRefine - Exercise (using data sample)

- We merge same entries in *president*
 - Trim
 - Titlecase
 - Text facet
 - Count
- We remove duplicates
 - Blank down or
 - Duplicates facet



o.	president	prior	party
	Facet	Text facet	District (1847-1849)
	Text filter	Numeric facet	see (1823-1825)
	Edit cells	Timeline facet	ates
	Edit column	Scatterplot facet...	(2005-2008)
	Transpose	Custom text facet...	a (1881-1887)
	Sort...	Custom numeric facet...	1979-1981 & 1983-1992)
	View	Customized facets	ates
	Reconcile	Word facet	Republican
		Duplicates facet	Offi
		Numeric log facet	
		1-bounded numeric log facet	
		Text length facet	
		Log of text length facet	
		Unicode char-code facet	
		Facet by error	
		Facet by null	
		Facet by empty string	
		Facet by blank (null or empty string)	



47 rows			
Show as: rows records Show: 5 10 25 50 100 500 1000 rows Sort ▾			
▼ All	▼ S.No.	▼ president	▼ prior
☆	16.	16	U.S. Representative for Illinois' 7th District (1847-1849)
☆	7.	7	U.S. Senator (Class 2) from Tennessee (1823-1825)
☆	17.	17	16th Vice President of the United States
☆	45.	44	s (2005-2008)
☆	23.	23	
☆	43.	42	
☆	31.	30	
☆	21.	21	
☆	47.	45	
☆	35.	34	
☆	33.	32	Franklin D. Roosevelt
☆	14.	14	Franklin Pierce
☆	42.	41	George H. W. Bush
☆	44.	43	George W. Bush
☆	1.	1	46th Governor of Texas (1995-2000)
☆	39.	38	George Washington
☆	22.	22	Commander-in-Chief of the Continental Army
☆	39.	38	40th Vice President of the United States
☆	22.	22	28th Governor of New York (1883-1891)
☆	24.	24	22nd President of the United States
☆	34.	33	34th Vice President of the United States

<https://guides.library.illinois.edu/openrefine/duplicates>

Text clustering

- We take a look at a text facet in *president*
- We can cluster groups of values that might be the same “thing”

The screenshot shows the OpenRefine web interface. At the top, the title bar reads "OpenRefine Processing and Cleansing Data 19-01-2024" with a "Permalink" button. Below this is a "Facet / Filter" section with "Undo / Redo 9 / 9" and a "45 rows" indicator. A "Refresh" button and "Reset all" / "Remove all" links are also present. The main data table shows a list of presidents with columns for "All", "S.No.", and "A". The "president" column is selected, and a "Cluster" button is visible. A modal window titled "Cluster and edit column 'president'" is open, displaying a table of clusters. The table has columns for "Cluster size", "Row count", "Values in cluster", "Merge?", and "New cell value". Two clusters are shown: one for "Barack Obama" and "Barak Obama", and another for "George H. W. Bush" and "George W. Bush". To the right of the modal, there are two sliders: "Average length of choices" (ranging from 11.5 to 15.5) and "Length variance of choices" (ranging from 0.5 to 1.5). The text "2 clusters found" is displayed at the top right of the modal.

OpenRefine Processing and Cleansing Data 19-01-2024 [Permalink](#)

Facet / Filter Undo / Redo 9 / 9 45 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

Refresh Reset all Remove all

president change

45 choices Sort by: name count Cluster

Abraham Lincoln 1
Andrew Jackson 1
Andrew Johnson 1
Barack Obama 1
Barak Obama 1
Benjamin Harrison 1
Bill Clinton 1
Calvin Coolidge 1
Chester A. Arthur 1
Donald Trump 1
Dwight D. Eisenhower 1
Franklin D. Roosevelt 1

Cluster and edit column "president"

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method Key collision Keying function Metaphone3 2 clusters found

Cluster size	Row count	Values in cluster	Merge?	New cell value
2	2	- Barack Obama - Barak Obama	☐	Barack Obama
2	2	- George H. W. Bush - George W. Bush	☐	George H. W. Bush

Average length of choices 11.5 — 15.5

Length variance of choices 0.5 — 1.5

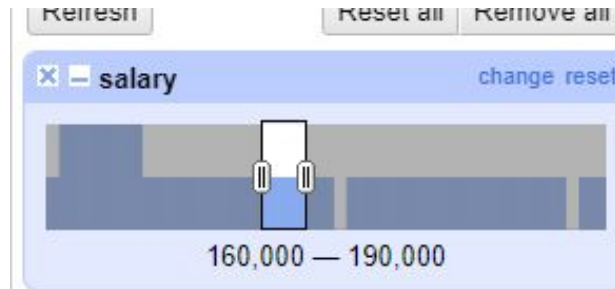
Numbers

- We take a look at a numeric facet in *salary*
 - Number Format: 111.11
 - → Replace . with nothing and , with .
 - Or GREL:
`value.replace(/[\$\.\.]/, "").replace(',','.').toNumber()`
 - Numeric Facet

▼ salary	▼ date updated	▼ date created
Facet	700:00:00Z	2012-04-03T00:00:00Z
Text filter	700:00:00Z	2012-04-03T00:00:00Z
Edit cells	700:00:00Z	2012-04-03T00:00:00Z
Edit column	700:00:00Z	2012-04-03T00:00:00Z
Transpose	700:00:00Z	2012-04-03T00:00:00Z
Sort...	700:00:00Z	2012-04-03T00:00:00Z
View	700:00:00Z	2012-04-03T00:00:00Z
Reconcile	700:00:00Z	2012-04-03T00:00:00Z
265000	2021-07-14	
75000	2021-07-14	
355000	2021-07-14	

▼ vice	▼ salary	▼ date updated	▼ date created
Hannibal Hamlin	Facet	700:00:00Z	2012-04-03T00:00:00Z
John C. Calhoun	Text filter	700:00:00Z	2012-04-03T00:00:00Z
Office vacant	Edit cells	700:00:00Z	2012-04-03T00:00:00Z
Joe Biden	Transform...	700:00:00Z	2012-04-03T00:00:00Z
Trim leading and trailing whitespace	Common transforms	700:00:00Z	2012-04-03T00:00:00Z
Collapse consecutive whitespace	Fill down	700:00:00Z	2012-04-03T00:00:00Z
Unescape HTML entities	Blank down	700:00:00Z	2012-04-03T00:00:00Z
Replace smart quotes with ASCII	Split multi-valued cells...	700:00:00Z	2012-04-03T00:00:00Z
To titlecase	Join multi-valued cells...	700:00:00Z	2012-04-03T00:00:00Z
To uppercase	Cluster and edit...	700:00:00Z	2012-04-03T00:00:00Z
To lowercase	Replace...	700:00:00Z	2012-04-03T00:00:00Z
To number	700:00:00Z	2012-04-03T00:00:00Z	2012-04-03T00:00:00Z
To date	700:00:00Z	2012-04-03T00:00:00Z	2012-04-03T00:00:00Z
To text	700:00:00Z	2012-04-03T00:00:00Z	2012-04-03T00:00:00Z
To null	700:00:00Z	2012-04-03T00:00:00Z	2012-04-03T00:00:00Z
To empty string	700:00:00Z	2012-04-03T00:00:00Z	2012-04-03T00:00:00Z

▼ salary	▼ date updated	▼ date created
Facet	Text facet	
Text filter	Numeric facet	
Edit cells	Timeline facet	
Edit column	Scatterplot facet...	
Transpose	Custom text facet...	
Sort...	Custom numeric facet...	
View	Customized facets	
Reconcile	700:00:00Z	2012-04-03T00:00:00Z
265000	2021-07-14T00:00:00Z	2012-04-03T00:00:00Z



Dates

- Transforms
 - toDate()
 - value.toDate('d.M.y')
 - value.toDate('E, d. MMM yyyy')
- Timeline Facet
- → <https://openrefine.org/docs/manual/grelfunctions#date-functions>

Custom text transform on column date updated

Expression Language General Refine Expression Language (GREL) ▼

`value.toDate('E, d. MMM yyyy')` No syntax error.

Preview History Starred Help

row	value	value.toDate('E, d. MMM yyyy')
16.	2021-07-14T00:00:00Z	[date 2021-07-14T00:00:00Z]
7.	2021-07-14T00:00:00Z	[date 2021-07-14T00:00:00Z]
17.	2021-07-14T00:00:00Z	[date 2021-07-14T00:00:00Z]
44.	Mittwoch, 14. Juli 2021	[date 2021-07-14T00:00:00Z]
23.	2021-07-14T00:00:00Z	[date 2021-07-14T00:00:00Z]
42.	2021-07-14T00:00:00Z	[date 2021-07-14T00:00:00Z]

On error ☒ keep original ☐ set to blank ☐ store error ☐ Re-transform up to times until no change

OK Cancel

General Refine Expression Language (GREL)

Common usage:

`value.toUpperCase(); value.toLowerCase(); value.titlecase(); value.trim()`

Date Formatting: `value.toString("dd. MMMM, yyyy")`

Strings: `value.contains("oe"), startsWith(), endsWith(), indexOf(), substring(), replace()`

Others: `value.type(), parseJson(), parseXml(), parseHtml()`

<https://openrefine.org/docs/manual/grelfunctions>

OpenRefine - Reconciliation

- reconciliation = alignment / matching
- connection to other databases is established
- additional information in these databases becomes available for OpenRefine
- Wikidata (en) is added by default
- additional services can be added
 - Wikidata German: <https://wikidata.reconci.link/de/api>
 - Description: <https://github.com/OpenRefine/OpenRefine/wiki/Reconcilable-Data-Sources>
 - **Overview:** <https://reconciliation-api.github.io/testbench/>
 - e.g. GND <https://lobid.org/gnd/reconcile>
 - Documentation: <https://openrefine.org/docs/manual/reconciling>
 - Framework for creating services: <https://github.com/codeforkjeff/conciliator>
 - API: <https://openrefine.org/docs/technical-reference/reconciliation-api>



- Information in structured / machine-readable format (RDF)
 - Data come FROM Wikipedia and go back TO Wikipedia
 - Community based, everyone can add concepts/items
 - Ontology is also community based (property proposals)
 - Goal: collect and connect common knowledge about the world (Wikipedia for machines) - language independent
 - <https://www.wikidata.org/wiki/Q23>
- Q Number = Item = concept and Object / Subject
 - Example
 - Q1 - the Universe
 - Q2 - Earth
 - Q5 - human
 - Q571 - book
 - Q7075 - library
 - P Number = Property = Prädikat
 - Example
 - P31 - instance of
 - P279 - subclass of
 - P214 - VIAF ID

OpenRefine Reconciliation Exercise

- column president
 - reconcile with Wikidata
 - add new columns with birth date and cause of death
- column vice
 - add GND service
 - reconcile
 - add a new column with the GND numbers

OpenRefine → Wikidata

1. query Wikidata for cities in AT without GND number:

https://query.wikidata.org/#SELECT%20%3Fitem%20%3FitemLabel%20WHERE%20%7B%0A%20%20%20%3Fitem%20wdt%3AP31%20wd%3AQ515%3B%0A%20%20%20%20%20%20%20%20%20%20wdt%3AP17%20wd%3AQ40.%0A%20%20%20FILTER%20NOT%20EXISTS%20%7B%20%3Fitem%20wdt%3AP227%20%3Fgnd%20%7D%0A%20%20SERVICE%20wikibase%3Alabel%20%7B%20bd%3AserviceParam%20wikibase%3Alanguage%20%22%5BAUTO_LANGUAGE%5D%2Cen%22.%20%7D%0A%7D%20LIMIT%2050

2. download the csv file and import to OpenRefine
3. reconcile “item” with wikidata and reconcile “itemLabel” with GND
4. add new column based on reconciled values from “itemLabel” with GND number
5. edit Wikidata Schema

Wikibase Schema

item

remove

Terms

Alias

▼

de

☐ override if present

remove

+ add term

Statements

GND ID

GND-ID

0 references

remove

+ add qualifier

+ add reference

+ add value

+ add statement

+ add item

OpenRefine Extensions

- <https://openrefine.org/extensions>
- AI
- RDF
- FAIR data
- GeoJSON
- Named-Entity Recognition
- ...

OpenRefine Export

Standard Methods: TSV, CSV, Excel, ODF, HTML

Custom tabular: Change Preferences and Upload to Google Spreadsheets ...

RDF Extension: RDF/XML and RDF/Turtle
Templating: customizable (JSON, XML ...)

Database Export: SQL

Export the entire project

Wikidata: Upload, QuickStatements, Schema

