

Large Language Models & Prompt Engineering Grundlagen

Angewandte Generative KI und Prompt Engineering
im Forschungsprojektmanagement
(Projektmanagement)

Dr. Christopher Pollin

<https://chpollin.github.io> | christopher.pollin@dhcraft.org

Digital Humanities Craft OG

www.dhcraft.org

Slides were generated AI-assisted.
Images are partly AI-generated.



Wie LLMs funktionieren

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

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

LLMs führen **Next Token Prediction** durch. Sie sagen das nächste Token in einer Folge von **Tokens** (~ **Kontext**) auf Grundlage ihrer **Trainingsdaten** voraus. Jedes vorhergesagte Token wird Teil des Kontexts für die nächste Vorhersage (**autoregressiv**). Dieser einfache Mechanismus, massiv skaliert, erzeugt die “schwach emergenten” (?) Eigenschaften, die wir beobachten.

Duchess = 98.55%

Duch = 1.28%

Du = 0.04%

Dutch = 0.03%

Duc = 0.02%

D = 0.01%

= 0.01%

Dou = 0.01%

Duke = 0.01%

\n\n = 0.01%

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

Andrej Karpathy. Deep Dive into LLMs like ChatGPT.

<https://youtu.be/7xTGNNLPyMI>

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

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

https://www.youtube.com/watch?v=zjkBMFhNi_g

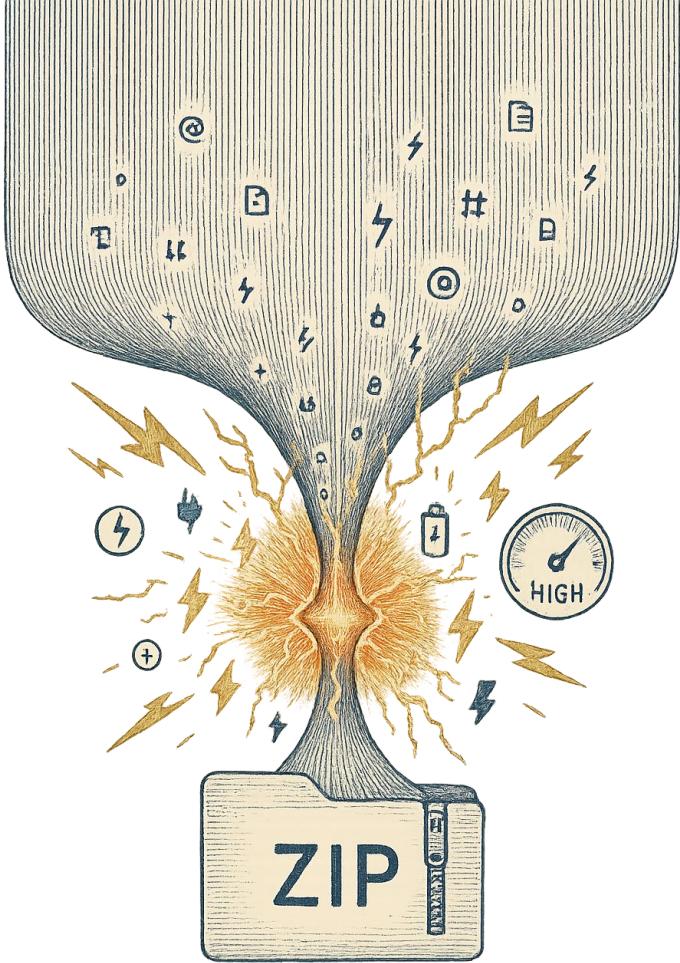
Alan Smith. Inside GPT – Large Language Models Demystified

<https://youtu.be/MznD2DzIQCe>

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>



Pre-Training (“Kompression von Wissen”)

- **Input:** Trillionen von Tokens aus (Web-)Daten und/oder synthetischen Daten
- **Task:** das nächste Token vorhersagen
- **Eigenschaften:**
 - verlustbehaftet (nicht perfekter Speicher)
 - probabilistisch (Muster, keine Fakten)
 - Wissensabgrenzung (zeitlich festgelegt)
- **Kosten** Sehr teuer (Geld, Energie, GPU), langsam

“Große Sprachmodelle sind verlustbehaftete, probabilistische Komprimierungen („.zip“) von möglichst vielen hochwertigen (multimodalen) Daten.”

Die *Gestalt* eines Wikipedia-Artikels über Zebras

Zebras ⋮A 151 Sprachen

Artikel Diskussion Lesen Bearbeiten Quelltext bearbeiten Versionsgeschichte Werkzeuge

(Weitergeleitet von [Zebra](#))

Dieser Artikel behandelt die Tiere – zu anderen Bedeutungen siehe [Zebra](#) (Begriffsklärung).

Die **Zebras** (*Hippotigris*) sind eine Untergattung aus der Gattung der **Pferde** (*Equus*), die aus drei Arten besteht: **Grevyzebra** (*Equus grevyi*), **Bergzebra** (*Equus zebra*) und **Steppenzebra** (*Equus quagga*). Die Tiere sind besonders durch ihr schwarz-weißes Streifenmuster gekennzeichnet. Alle Vertreter der Zebras kommen ausschließlich in Afrika vor, wo sie zumeist offene Landschaften bewohnen. Die Bestände der einzelnen Arten sind in unterschiedlichem Maße bedroht.

Name [Bearbeiten] [Quelltext bearbeiten]

Als [portugiesische Seefahrer](#) Ende des 15. Jahrhunderts erstmals Zebras bei ihren Entdeckungsfahrten in Afrika sahen, fühlten sie sich an die [iberischen Wildpferde](#) erinnert, die sie in [Portugal](#) **Zebras** nannten und die ebenfalls schwarze Streifen zeigten. Fortan nannten sie diese Pferde **Zebras**.^[1]

Zebras

Steppenzebra (*Equus quagga*)

Systematik

Unterkategorie: Höhere Säugetiere

LLMs können nicht direkt auf Wikipedia-Artikel zugreifen. Sie haben nur Zugriff auf die **Gestalt** (Karpathy) des Textes, die komprimierte statistische Muster darstellt, die während des Trainings entstehen.

LLMs besuchen Webseiten nicht! Sie können jedoch Tools für die Websuche verwenden (**Tool Use**).

Interne Wissensrepräsentation des Modells im Vergleich zu seiner Fähigkeit, über Tools auf externe Informationen zuzugreifen

Character Claude

<https://www.anthropic.com/research/clause-character>

<https://www.youtube.com/watch?v=ugvHCXCOmm4&t=9774s>



Meta Builds Manhattan-Sized AI Data Centers in Multi-Billion Dollar Tech Race.

<https://www.ctol.digital/news/meta-builds-manhattan-sized-ai-data-centers-tech-race/>



Inside OpenAI's Stargate Megafactory with Sam Altman | The Circuit.

<https://youtu.be/GhJJs4zbH0o>

Die USA investieren Hunderte von Milliarden in Rechenzentren und Energieerzeugung.

- **Energie:** 0,0003 kWh pro Prompt (= 8–10 Sekunden Netflix-Streaming)
- **Wasser:** 0.25-5mL pro Prompt (feinige Tropfen bis zu 1/5 eines Schnapsglases)
- **Effizienz:** 33-fache Verbesserung in einem Jahr (Google)
- **Kosten:** \$50→\$0.14 per million tokens (GPT-4 to GPT-5 nano)

Ethan Mollick. Mass Intelligence. From GPT-5 to nano banana: everyone is getting access to powerful AI <https://www.oneusefulthing.org/p/mass-intelligence>

Während einzelne LLM-Abfragen immer effizienter werden, führt ihr massiver Einsatz zu einem Paradoxon, bei dem allein GPT-4o jährlich so viel Strom verbraucht wie 35.000 US-Haushalte. Dies zeigt, dass die Wahl der Infrastruktur für die Umweltbelastung wichtiger ist als die Modellgröße und dass die weltweite Einführung von KI einen Ressourcenverbrauch verursacht, der die Effizienzgewinne bei weitem übersteigt.

Jegham, Nidhal, Marwen Abdelatti, Lassad Elmoubarki, and Abdeltawab Hendawi. 'How Hungry Is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference'. 14 May 2025. <https://doi.org/10.48550/arXiv.2505.09598>.



<https://huggingface.co/datasets/HuggingFaceFW/fineweb>

Tokens	Characters
3	12
Hello World!	

<https://platform.openai.com/tokenizer>
<https://tiktoktokenizer.vercel.app/>

Tokenization

- **Rohtext aus dem Internet:**
 - “Hello World!”
- **Cleaning und Filtering**
 - (removes spam, deduplication)
- **Tokenizer**
 - [‘Hello’, ‘World’, ‘!’]
- **IDs**
 - [13225, 5922, 0]

Die Tokenisierung wandelt Text in numerische Einheiten für die LLM-Verarbeitung um. Die Tokenisierungsstrategie priorisiert die Recheneffizienz, indem sie die Sequenzlänge minimiert.

Ein Token ist die atomare Einheit für LLMs (100 Token ≈ 75 englische Wörter).

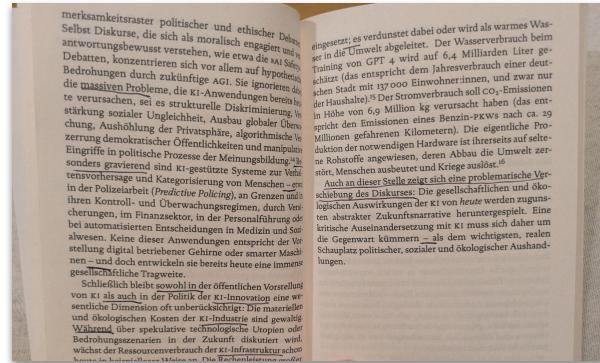
Deep Dive into LLMs like ChatGPT.
<https://youtu.be/7xTGNNLPyM>
Let's build the GPT Tokenizer.
<https://youtu.be/zduSFxRajkE>

Warum kann ein LLM nicht buchstabieren und warum sieht man so viele „:“ und „—“?

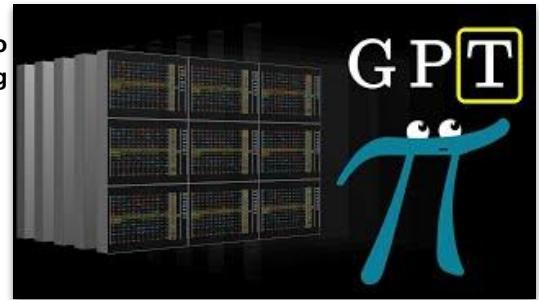
The image shows a screenshot of a ChatGPT 5 interface. At the top left, it says "ChatGPT 5". At the top right, there is an upward arrow icon and the word "Get". In the center, there is a light gray rounded rectangle containing the following text:
How many Bs are there in blueberry?
Give the answer directly!
Below this box, the number "3" is displayed. At the bottom of the screen, there are several small icons: a square with a dot, a downward arrow, a checkmark, a double arrow, a double downward arrow, a double upward arrow, a double checkmark, and a circular arrow.

Joiner	Token sequence (conceptual)	Token count
" __ "	[" — "]	1
", and "	[" , ", " and"]	2

Let's talk about em dashes in AI. Maria Sukhareva. <https://msukhareva.substack.com/p/lets-talk-about-em-dashes-in-ai>



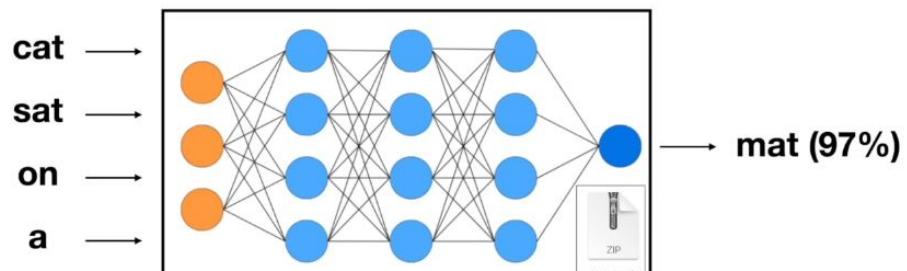
AI Slop



Transformer-Architecture

Neural Network

Predicts the next word in the sequence.



e.g. context of 4 words

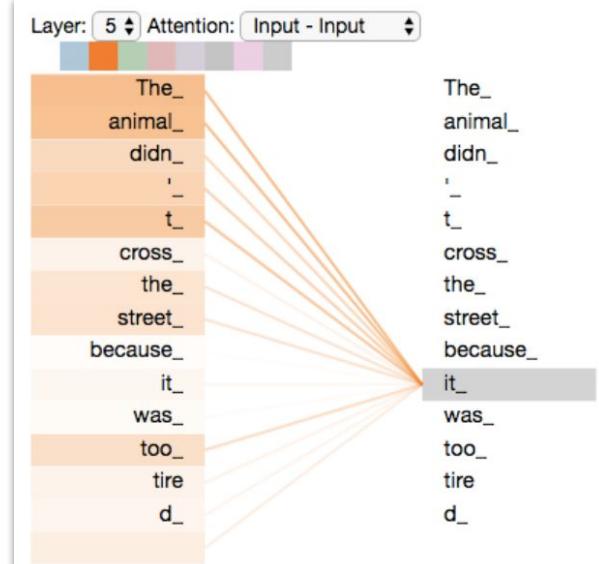
predict next word

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, 2024

<https://youtu.be/MznD2DzIQCc>



<https://jalammar.github.io/illustrated-transformer>

Model Context Window = 8K

Input Token

6000 Token

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

[...]

Output Token

1500 Token

Lorem ipsum ...

Context Window = $6000 + 1500 < 8000$

Model Context Window = 8K

10000 Token

A context window, in the context of large 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.

[...]

1500 Token

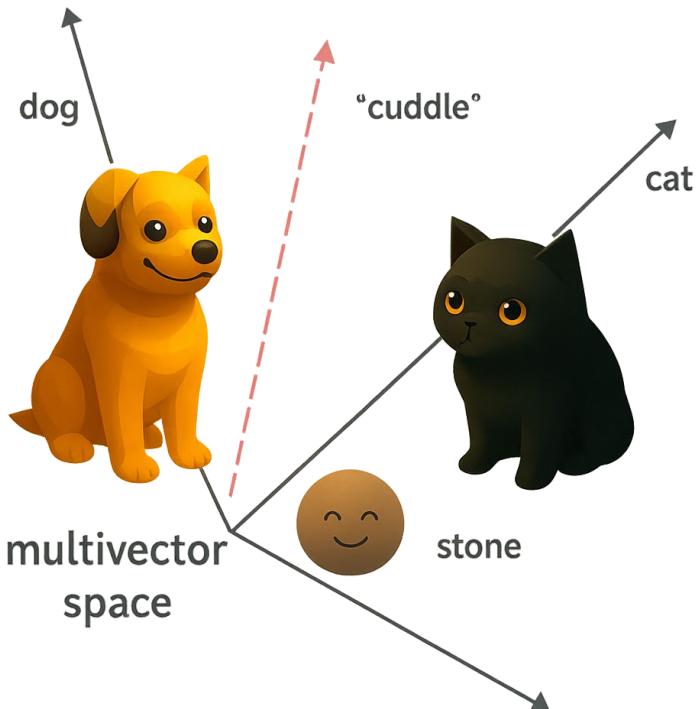
Lorem ipsum ...

Context Window = $10000 + 1500 > 8000$

3500 tokens are not in the context window!

Dingemanse, Mark. 2024. „Generative AI and Research Integrity“. Preprint, OSF, Mai 14. <https://doi.org/10.31219/osf.io/2c48n>.

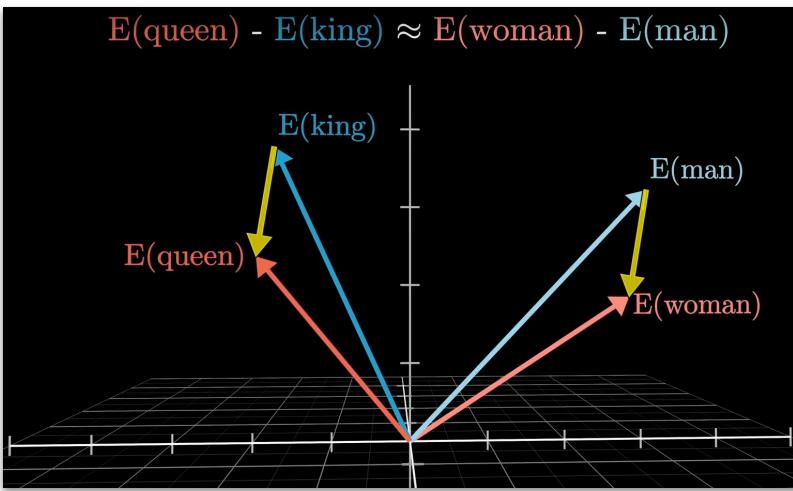
Embeddings



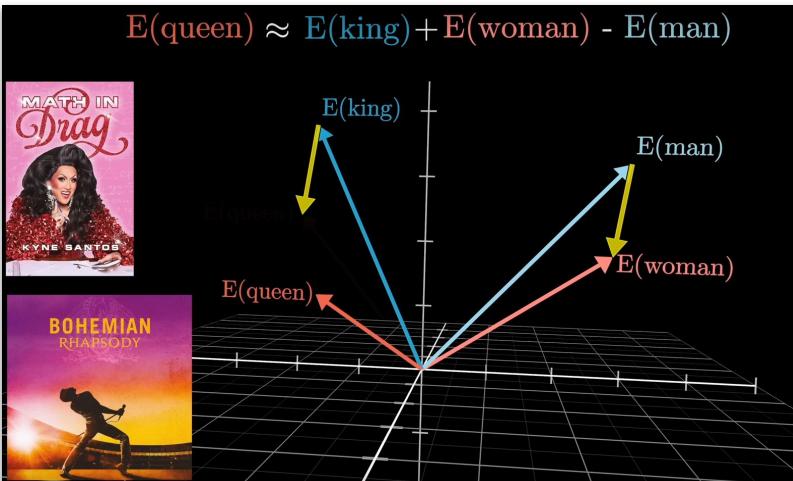
- **Ähnliche Bedeutungen = Nächere Positionen im Raum**
 - “dog” und “cat” → nah
(beides Haustiere, Tiere, Säugetiere)
 - “stone” → Fern (lebloses Objekt)
 - “cuddle” → Näher an Tieren
(Handlung, die mit Lebewesen assoziiert wird)
- **Multivektorraum**
 - n Dimensionen (GPT 3.5 ~ 15.000)
 - Positionen entstehen aus dem Pre Training

Einbettungen wandeln diskrete Token (Wörter) in kontinuierliche numerische Vektoren in einem hochdimensionalen Raum um.

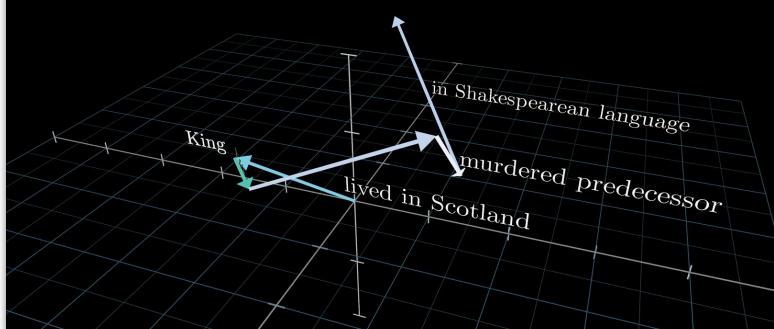
Embeddings



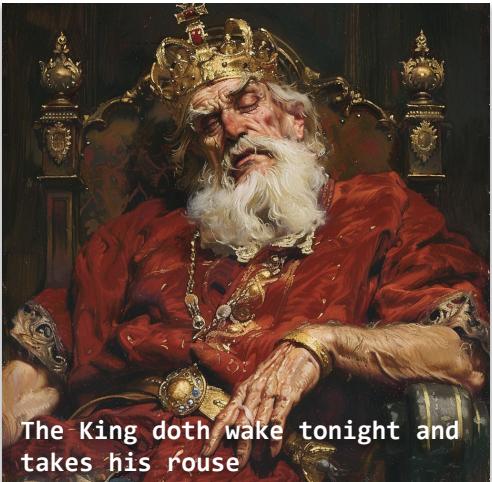
[3Blue1Brown. But what is a GPT? Visual intro to transformers | Chapter 5, Deep Learning](#)



The King doth wake tonight and takes his rouse



“Shakespearean English”



You

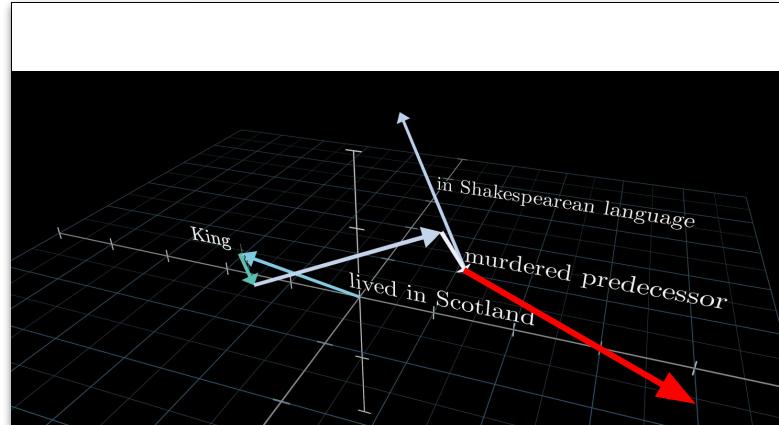
The King doth wake tonight and takes his rouse

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

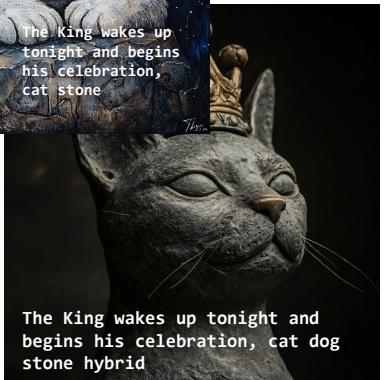
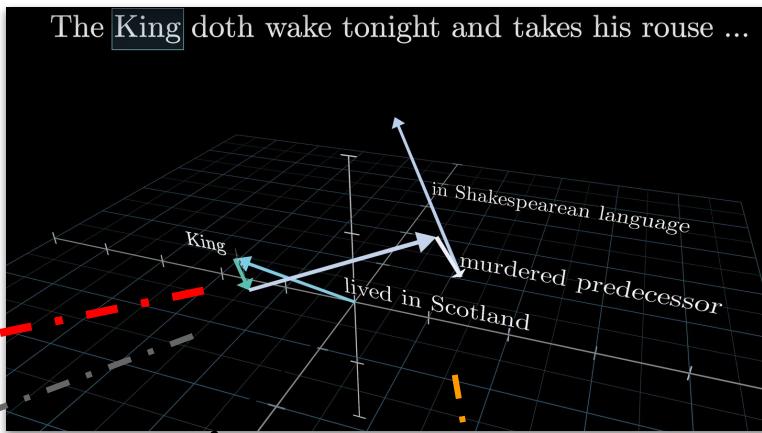


“Modern Englisch”

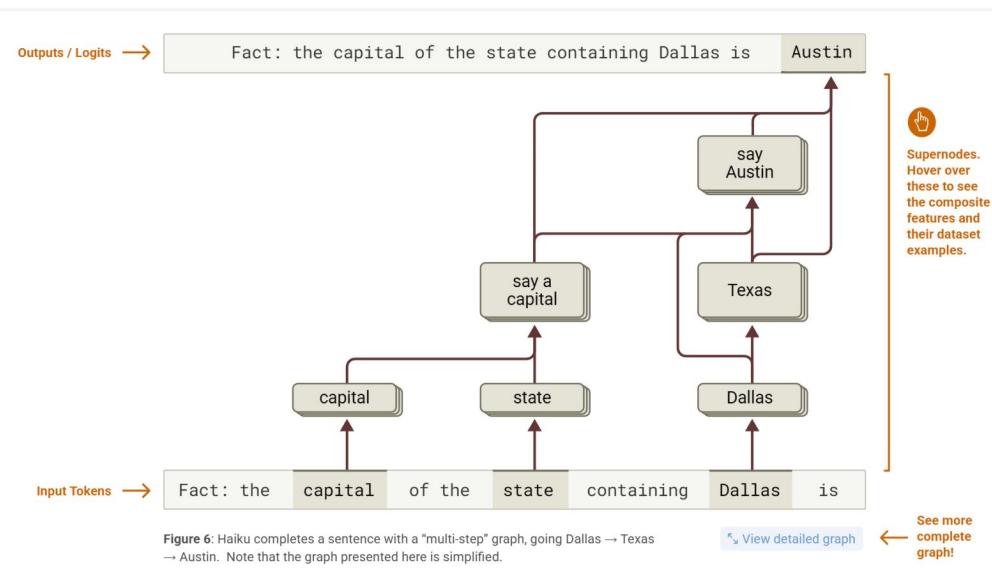




cat
dog
stone
hybrid

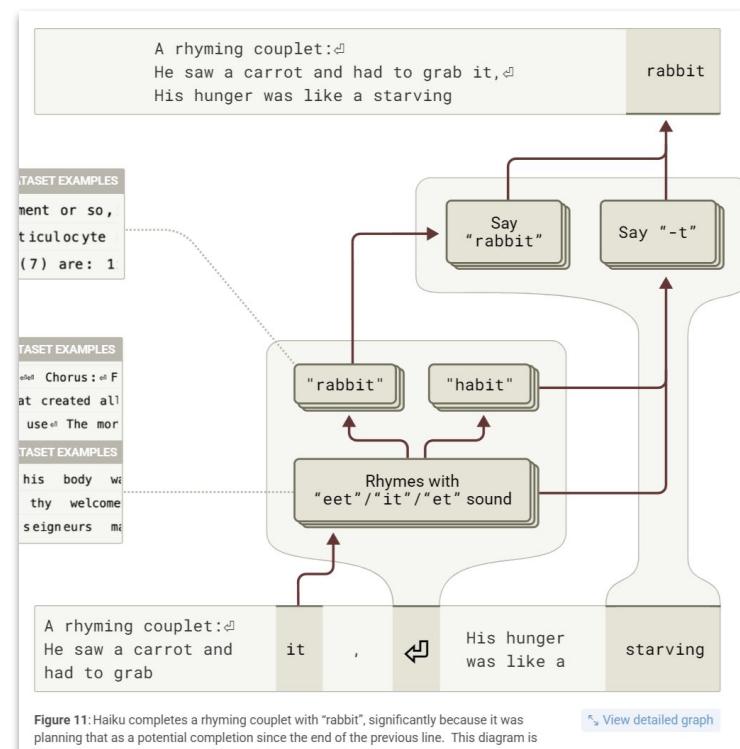


“On the Biology of a Large Language Model”



Anthropische Forscher haben die internen Berechnungen von Claude 3.5 Haiku nachverfolgt. Sie fanden echte mehrstufige „Argumentationen“ (Dallas → Texas → Austin), vorausschauende Planung in der Poesie (Aktivierung von Reimwörtern vor dem Schreiben von Zeilen) und sprachunabhängige Darstellungen in Englisch, Französisch und Chinesisch.

Das Modell führt diese Berechnungen automatisch durch, ohne metakognitives Bewusstsein für seine eigenen Prozesse.



Anthropic. Interpretability: Understanding how AI models think. <https://youtu.be/fGKNUvivnc>
Anthropic. On the Biology of a Large Language Model.
<https://transformer-circuits.pub/2025/attribution-graphs/biology.html>

LLMs as Retrieval-*ish* Systems

“LLMs are stores of knowledge and programs - they've stored pattern from the internet as vector programs” (François Chollet)

LLMs rufen memorierte “Programme” aus dem latenten Raum ab und interpolieren zwischen ihnen, können aber nicht von gelernten Mustern abweichen. Ihre lückenhafte Generalisierung versagt bei Unbekanntem. Prompt Engineering sucht die optimalen Koordinaten für diese Programme.

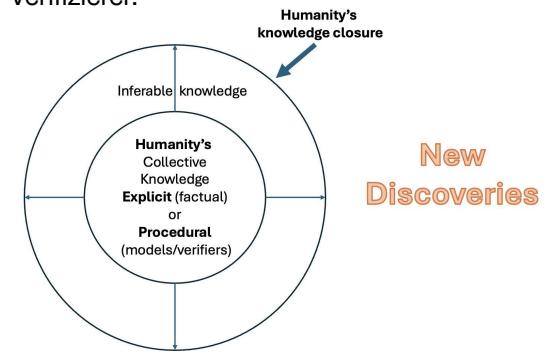
“Large language models is for me a database technology. It's not artificial intelligence.”
(Sepp Hochreiter)

LLMs erfassen menschliches Wissen aus Text/Code und speichern es. Aktuelles “Reasoning” wiederholt nur bereits gesehene Denkmuster. Sie können keine genuinen neuen Konzepte oder Denkansätze erschaffen.

xLSTM wird als Alternative entwickelt.

“LLMs are n-gram models on steroids doing approximate retrieval, not reasoning”
(Subbarao Kambhampati)

Approximatives Retrieval täuscht Reasoning durch Mustererkennung vor, versagt aber bei Verschleierung und benötigt externe Verifizierer.





Prompt Engineering: “How to Prompt a Research Blog”

Properitäre vs. Open Source vs. Open Weights LLM

Claude Opus 4.1 | Claude Sonnet 4.5

<https://claude.ai>

Gemini 2.5 Pro

<http://gemini.google.com> |
<https://aistudio.google.com>

GPT-5

<https://chatgpt.com>

Qwen3-Max

<https://qwen.ai>

Mistral

<https://chat.mistral.ai>

Apertus

<https://www.swiss-ai.org/apertus>

Aleph Alpha

<https://aleph-alpha.com>

...

ChatGPT ≠ AI ≠ GPT ≠ Machine Learning ≠
LLM ≠ Open Source ≠ Open Weights ≠ ...

Wo LLM Probleme haben

Rechnen

Buchstabieren, Zählen

Websuche

Aktuelle Informationen

Räumliches und Zeitliches “Denken”

“Denken”

Stil (Sycophancy)

Nicht deterministisch, d. h. es kommen immer andere Antworten.

Es gibt keinen Mechanismus in einem LLM, um etwas als “wahr” zu verifizieren → “Expert-in-the-Loop”

Assignment 1: Forschungsblog auf GitHub Pages

Deadline: 17.11.2025, 23:59 Uhr

Aufgabenstellung: Erweitere dein Projektkonzept aus Assignment 0 (Pitch) zu einem digitalen Forschungsblog und publiziere diesen über GitHub Pages. Der Blog soll deine Projektidee vertiefen und einen spezifischen Aspekt deines Vorhabens detailliert beleuchten. **Nutze LLMs aktiv für die Texterstellung und die Entwicklung der Webseite** – dokumentiere in deinem Prompt Engineering Journal, welche Prompts und Tools du verwendet hast und reflektiere den Einsatz!

- Technische Anforderungen:
 - Static HTML Website (mindestens index.html)
 - CSS Stylesheet (extern oder embedded)
 - Publikation via GitHub Pages
 - Projektkürzel (=Repo Name)
 - Mindestens 3 Commits mit sinnvollen Commit-Messages
 - Responsive Design (mobile-friendly)

Assignment 1: Forschungsblog auf GitHub Pages

Inhaltliche Anforderungen:

- Qualität > Länge
- Beispiel: <https://dhcraft.org/excellence/blog/System1-42>
- **Projektübersicht (erweitert aus 1-Pager)**
 - Forschungsfrage und Kontext
 - Datenbeschreibung mit Beispielen
 - Methodischer Ansatz
- **Vertiefender Fokus (wählen EINEN Aspekt):**
 - Datenaufbereitung: Herausforderungen und Lösungsansätze
 - Prompt Engineering: Geplante LLM-Interaktionen
 - Visualisierungskonzept: Darstellung der Ergebnisse
 - Forschungsrelevanz: Einbettung in Fachdiskurs
- **Mini-Projektplan**
 - Meilensteine
 - Technologie-Stack
 - Erwartete Herausforderungen

GitHub & GitHub Pages

Was ist GitHub?

- Versionskontrollplattform für Code und Texte
- Basiert auf Git (Versionskontrollsyste)
- Speichert alle Änderungen deines Projekts (Timeline)
- Kollaboration: Mehrere Personen können am gleichen Projekt arbeiten
- Repository = Projektordner in der Cloud

Was ist GitHub Pages?

- Kostenloser Webhosting-Service von GitHub
- Verwandelt dein Repository in eine Website
- URL-Format: `username.github.io/repository-name`
- Perfekt für statische Websites (HTML, CSS, JavaScript)
- Automatische Publikation bei jedem Push

Setup in 4 Schritten

1. **GitHub Account erstellen**
 - a. Registrierung auf github.com
2. **GitHub Desktop installieren:**
<https://desktop.github.com/download>
 - a. Grafische Oberfläche für Git
 - b. Einfaches Commit & Push ohne Terminal
3. **Repository erstellen & klonen**
 - a. "New Repository" auf GitHub
 - b. Mit GitHub Desktop klonen
 - c. Lokal arbeiten, Änderungen committen & pushen
4. **GitHub Pages aktivieren**
 - a. Repository → Settings → Pages
 - b. Source: "Deploy from a branch"
 - c. Branch: main → / (root)
 - d. Nach 2-3 Minuten online unter `username.github.io/repository-name`

GitHub Pages

The screenshot shows the GitHub Pages settings page for a repository. The left sidebar lists various GitHub features: General, Access, Collaborators, Moderation options, Code and automation (Branches, Tags, Rules, Actions, Models), Webhooks, Copilot, Environments, Codespaces, and Pages (which is selected). The main content area is titled "GitHub Pages" and contains the following information:

- Access:** GitHub Pages is designed to host your personal, organization, or project pages from a GitHub repository.
- Build and deployment:**
 - Source:** Deploy from a branch (selected)
 - Branch:** Your GitHub Pages site is currently being built from the `/docs` folder in the `main` branch. [Learn more about configuring the publishing source for your site.](#)
 - Buttons: `main`, `/docs`, `Save`
- Custom domain:** Custom domains allow you to serve your site from a domain other than `chpollin.github.io`. [Learn more about configuring custom domains.](#)

At the bottom, there are "Save" and "Remove" buttons.

Hands-On: 1. Version auf GitHub Pages bringen

Verwende idealerweise ein Frontier-LLM und das kommt alles in deinen Ordner fürs Repo:

- Akademisches Narrativ für einen Forschungsblog extrahieren und bearbeiten
- Kümmere dich nicht so sehr um Formulierungen. Achte vielmehr auf das Narrativ.
- Baue 1-2 Quellen ein (Paper, Daten, andere Ressourcen)
- Lege ein ganz kleines [DESIGN.md](#) an (Design)
- Lege ein ganz kleines [RESEARCH-BLOG.md](#) an (Inhalt)
- Erzeuge eine einfache erste Webseite
 - index.html
 - style.css
 - Oder ein Python Script dass aus dem [RESEARCH-BLOG.md](#) den Text lädt und ein HTML erzeugt.
- Dokumentiere in einem Prompt Engineering Journal und mach auch sehr saubere Commits.

Hands-on: Extraktion des akademischen Narrativs

Manfred Thaller: Can historical information be displayed outside of a graph/hypergraph/network?:

{copy/paste}

Analyse the text in detail and step by step and extract the academic narrative as list.

Manfred Thaller: Can historical information be displayed outside of a graph/hypergraph/network?
<https://graphentechnologien.hypotheses.org/files/2021/02/Thaller-Mainz2021-2.pdf>

Trennzeichen und Struktur

Kontextinformation

Arbeitsstil

Instruktionen

Art der Antwort definieren

Laden Sie keine PDF-/Word-Dateien direkt hoch, sondern kopieren Sie den Text und fügen Sie ihn ein (meistens besser)

Die Reihenfolge der Anweisungen ist (ein bisschen) wichtig.

Follow-Up Prompts

Depending on the model used, you may need to rephrase these prompts to “find the better ‘programmes’”.

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



Write a detailed and concise report with ALL findings.



Can the same information be presented more concisely without losing any of its complexity or detail?

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



Write the perfect report with ALL information!

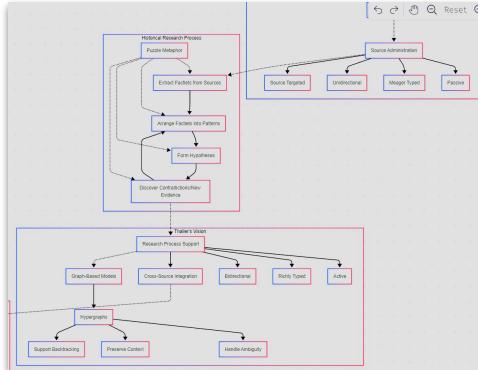
Think step by step about how to create a slide to give an overview and teach the content.

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 based on what appears in print"
- Factlet**: bits 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 are not documents that tokens relevant to various questions
- Meaning 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	Unidirectional	Knowledge gained isn't fed back into database
Databases don't track how their data is used or notify users of changes	Source Targeted	Source Unstructured
Meager typed: Semantic richness is lost when data is extracted	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>