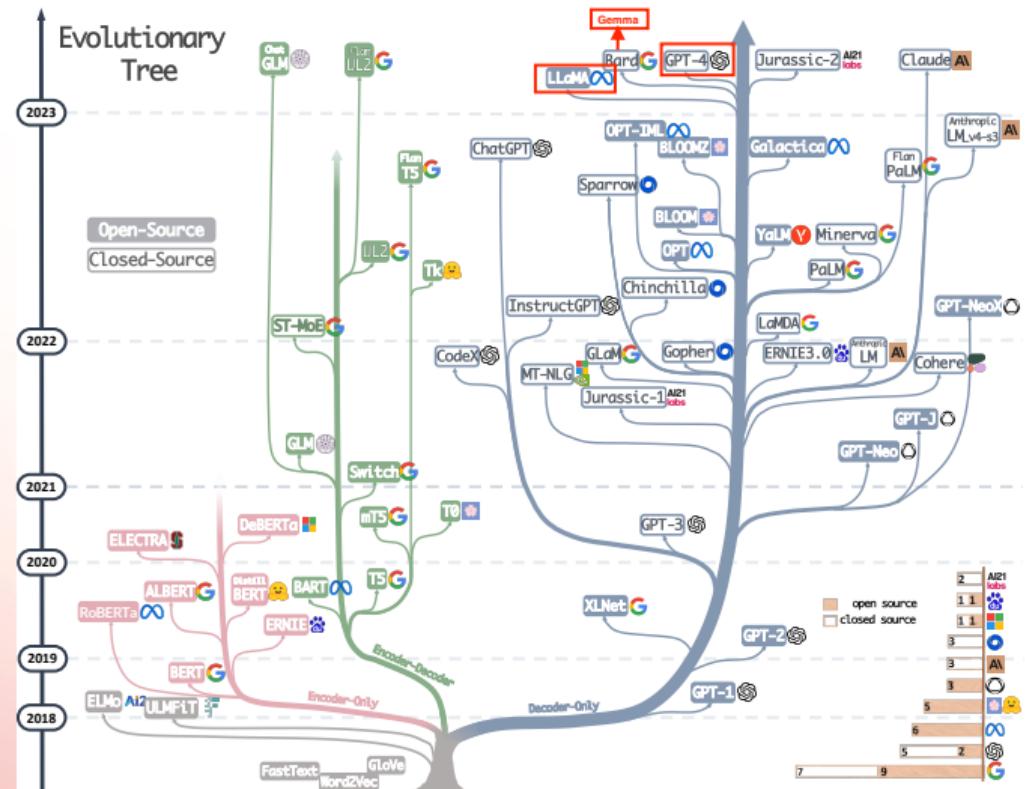


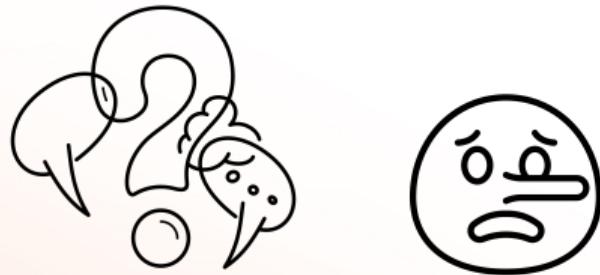
LLMs - Eine Einführung in RAG und Fine-Tuning

The Evolutionary Tree of Modern LLMs

arXiv: 2304.13712



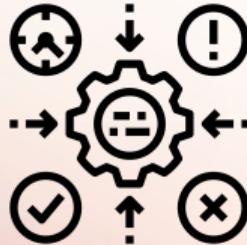
LLMs als Informationsquelle?



- LLMs generieren Text der sich “richtig anhört” (so kommt es auch zu vielen richtigen Antworten), sind aber keine Wissensdatenbank
- oft “halluzinieren” LLMs

Frage: Wie kann man LLMs dennoch nutzen, um Informationen z.B. aus einem Sachtext zu extrahieren?

LLMs als Informationsquelle!



1. Methode: **RAG** (Retrieval Augmented Generation): Füge externes Hintergrundwissen (sog. Kontext) zum Prompt hinzu



2. Methode: **Fine-Tuning**: einfügen des Hintergrundwissens in das LLM selbst.

- Ausgangspunkt: ein offene verfügbares, trainiertes LLM.
- Nachtraining des Modells mit den eigenen Daten

RAG - Retrieval Augmented Generation

Was kommt in den Kontext?



Naiver Ansatz: Alles. **Probleme:**

- evtl. passt nicht alles in den Kontext
- jede Anfrage ist sehr teuer und langsam

Beispiel: Eine Firma will z.B. Konstruktionsrichtlinien über einen Chatbot zugänglich machen.
Aber: Die maximale Promplänge des LLMs lässt nur maximal eine Richtlinie zu.

Ideen?

Was kommt in den Kontext?

Dies ist abhängig von der gestellten Frage!

- 1. Ansatz:** Finde zur Frage die passendste Richtlinie. Aber: So können keine dokumentenübergreifenden Informationen ausgegeben werden.



- 2. Ansatz:** Unterteile alle Richtlinien in z.B. Kapitel oder Abschnitte (**Chunking**) und finde zur Frage die besten Abschnitte.

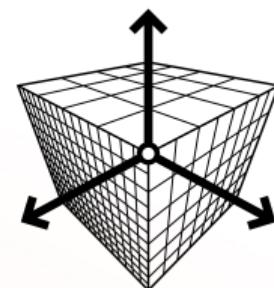


Frage: Wie finden wir die besten Abschnitte?

Information Retrieval durch semantische Ähnlichkeit

Hypothese: Die besten Abschnitte sind die, die der Frage semantisch am ähnlichsten sind.

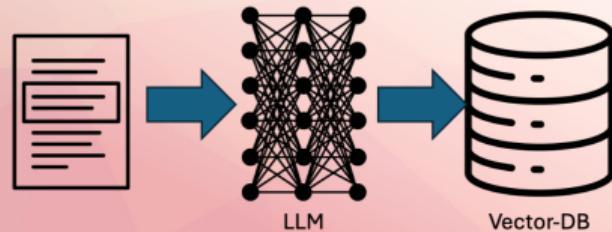
Fakt: LLM Encoder stellen Textabschnitte als Punkte in einen höherdimensionalen Vektorraum dar (=Embedding/Vektorisierung), wobei semantisch ähnlicher Text näher aneinander liegt.



Hypothese neu: Die besten Abschnitte sind die, die der Frage in dem Vektorraum am nächsten liegen.

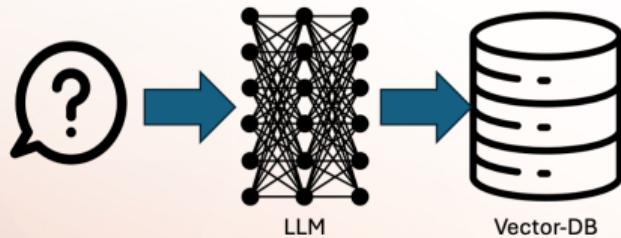
1. Schritt: Vektorisieren = Kontext einbetten

Chunking der “Hintergrundwissen”-Dokumente



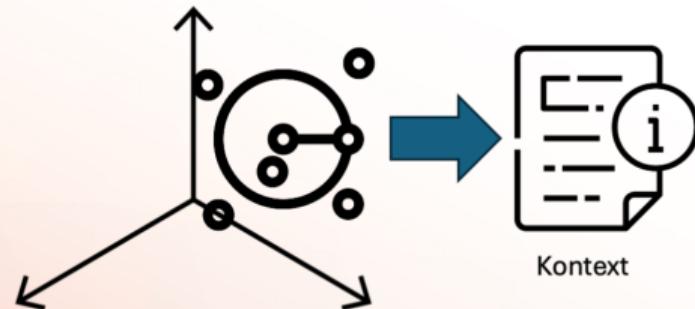
Vektorisieren der Abschnitte in die Vektordatenbank mittels eines LLM

2. Schritt: Retrieval



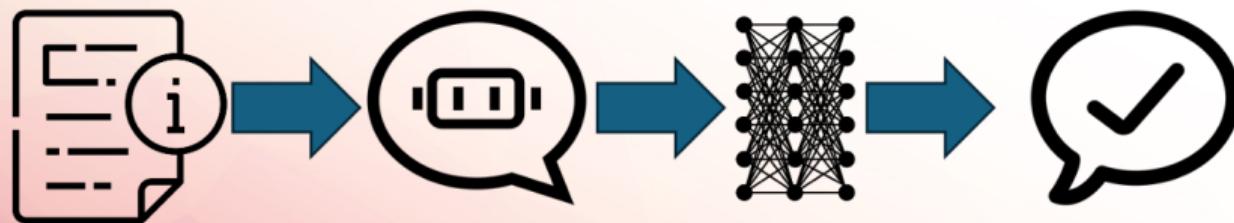
Embedding der Frage in die Vektordatenbank mittels eines LLM

Retrieval der k nächsten Punkte in der Vektordatenbank = Kontext

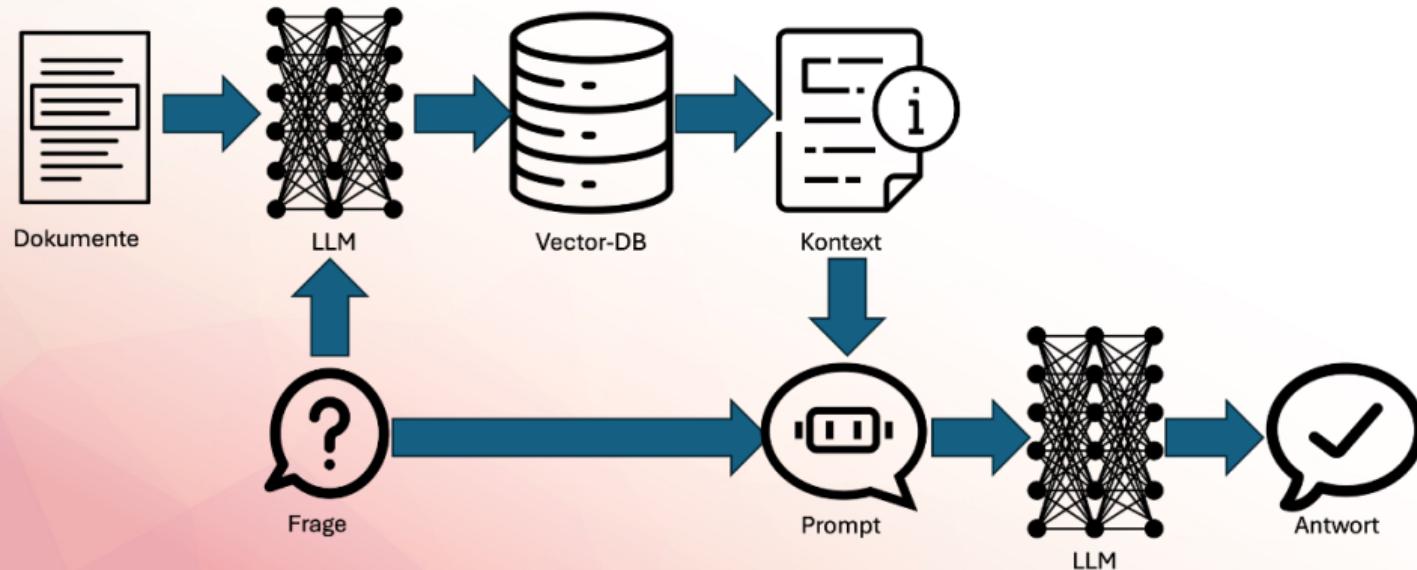


3. Schritt: Generieren einer kontext-basierten Antwort

Einfügen des Kontexts in einen vorgefertigten **Prompt** an ein **LLM** zum Generieren der Antwort.



Retrieval Augmented Generation (RAG) - Überblick



Was ist zu beachten?

- **Chunking:** Wie zerschneidet man die Dokumente am besten?
- **Retrieval:** Wie viele Abschnitte gehören in den Kontext? In welcher Reihenfolge? Weiteres Pre- or Post-Processing zum Sicherstellen der Relevanz?
- **Prompting:** Wie schreibt man den Prompt für beste Ergebnisse und “den richtigen Tonfall”?
- **Zusätzliche Methoden:** Fine-Tuning...

Dafür braucht es **Metriken** zur Beurteilung der Qualität.



Bewertung von RAG-Chatbots

Wie bewertet man RAG?

Hat man eine Frage, so liefert RAG zwei Outputs:

- den gefundenen **Kontext**
- die aus diesem Kontext generierte **Antwort**

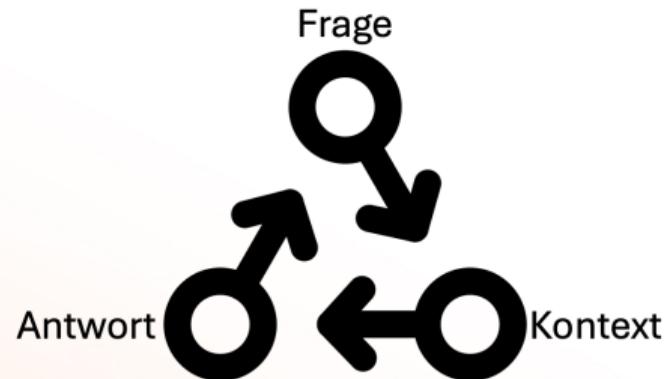
Was führt zu einer guten Antwort?

- “guter” Kontext (passend zur Frage und vollständig aber wenig Überflüssiges)
- aus dem Kontext wird eine “gute” Antwort generiert

Bewertungsgrundlage von RAG

RAG-generierte Q&A-Tripel

- **Frage**, die mit den Daten beantwortet werden kann
- vom RAG gefundener **Kontext** zur Frage
- vom RAG generierte **Antwort** auf Basis des Kontexts



Wie wird bewertet?

Frage: Wer sagt ob ein Q&A-Tripel “gut” ist?

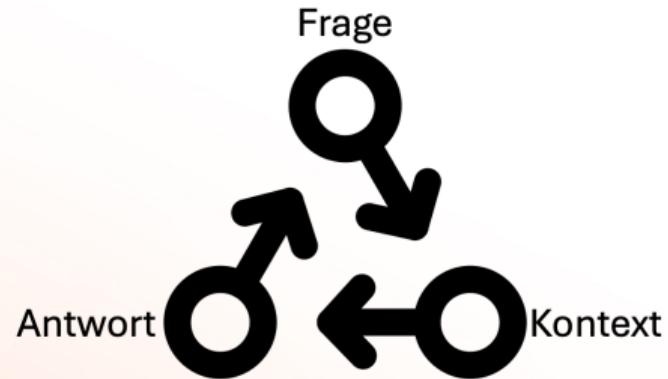
Zwei Ansätze:

- **ohne Ground Truth:** ein LLM (z.B. ChatGPT) bewertet via geeignetem Prompting wie gut alles zueinander passt
- **mit Ground Truth:** von Menschen vorgegebene oder von LLMs erstellte “richtige” Antworten werden (auch von einem LLM) mit den generierten Q&A-Tripeln verglichen.

Ground Truth als Referenzpunkt

Ground Truth: richtige Q&A-Tripel

- **Frage**, die mit den Daten beantwortet werden kann
- der richtige **Kontext** zur Frage
- die richtige **Antwort** auf Basis des Kontexts



Wie kommt man an Ground Truth?

Ground Truth = richtige Frage-Kontext-Antwort Tripel.

- **1. Möglichkeit:** per Hand erstellen (Vorteil: beste Qualität)
- **2. Möglichkeit:** durch LLM (ChatGPT) erstellen lassen (Vorteil: automatisierbar, man kann Massen generieren): Prompt = Kontext + die Anweisung, aus dem Kontext ein paar Fragen mit Antworten zu generieren (z.B. mit Few-Shot Prompting, siehe später)

Bewertungsmetriken ohne Ground Truth

Die folgenden Fragen können auch durch ein LLM beantwortet werden:

- **Context relevance:** Ist der gefundene Kontext relevant für die Frage?
- **Faithfulness:** Bezieht sich die Antwort vollständig auf den gefundenen Kontext?
- **Answer relevance:** Ist die Antwort relevant für die Frage?



Weitere Bewertungsmetriken für den gefundenen Kontext

Die folgenden Metriken brauchen jedoch Ground Truth:

- **Context recall:** Welcher Prozentsatz der Sätze aus der Ground Truth Antwort können aus dem gefundenen Kontext entnommen werden?
- **Context precision:** bewertet, inwieweit die Ground-Truth relevanten Aussagen unter den Kontext Chunks als wichtig gerankt werden.

End-to-End Bewertungsmetriken

- **Answer semantic similarity:** Wie ähnlich ist die Antwort einer vorgegebenen richtigen (= ground truth) **semantisch** auf einer Skala von 0 bis 1? (größer ist besser)
- **Answer factual similarity:** ... **inhaltlich** ...
- **Answer correctness:** ... **semantisch und inhaltlich**...

Weitere Details zu allen Metriken:

<https://docs.ragas.io/en/stable/concepts/metrics/index.html>

End-to-End Bewertungsmetriken: Formeln

Answer semantic similarity: Vektorisiere die korrekte Antwort sowie die generierte Antwort und berechne ihre **Cosine Similarity** (= Kosinus des Winkels zwischen beiden Vektoren)

Answer factual similarity:

- TP (True Positives): Aussagen die in beiden Antworten vorkommen
- FP (False Positives): Aussagen die in der generierten Antwort vorkommen, aber nicht in der richtigen
- FN (False Negatives): Aussagen die nicht in der generierten Antwort vorkommen, aber in der richtigen

Dann ist die factual similarity der entsprechende F1 Score = $\frac{TP}{TP + \frac{FP+FN}{2}}$

Answer correctness: (gewichteter) Mittelwert zwischen semantic similarity und factual similarity

RAGAS: Bewertung durch Prompting

Voraussetzung: Dataset an RAG-generierten Q&A-Tripeln.

Formuliere Prompts für jede Metrik, deren Antwort die Bewertung eines Q&A-Tripels ist (z.B. 0 oder 1). Alle Bewertungen werden aggregiert, z.B. via

$$\frac{\text{Anzahl von positive Antworten}}{\text{Anzahl von Tripeln}}$$



Alternative zu RAGAS: ARES.

ARES - Bewertung durch drei LLM-Judges

Voraussetzung: mindestens 150 von Menschen generierte Ground Truth Q&A-Tripel.

Grundidee: Nutze drei “kleine” selbst-trainierte LLMs als “Richter” über Frage-Kontext-Antwort-Tripel bezüglich je einer Metrik.



Die drei Schritte von ARES

- **1. Schritt:** benutze ein LLM um aus den Kontext-Passagen synthetische Fragen-Antwort Paare für alle drei Metriken zu generieren (sowohl korrekte (positiv), als auch falsche (negativ), zu gleichen Teilen)
- **2. Schritt:** Trainiere drei separate kleine LLMs anhand der synthetischen Fragen aus 1. darauf, ein Frage-Kontext-Antwort-Tripel als positiv oder negativ zu klassifizieren. Fine-tuning mit den 150 “handgelabelten” Tripeln.
- **3. Schritt:** Benutze die drei LLMs, um echte Frage-Kontext-Antwort-Tripel des RAGs zu klassifizieren, und berechne die Metriken wieder via

$$\frac{\text{Anzahl von Positiven}}{\text{Anzahl von Tripeln}}$$

RAG Performance-Optimierung

Meta-Data Filtering, Query Rephrasing, Reranking

Meta-Data Filtering: Filtern der Kontext-Abschnitte nach Meta-Daten
⇒ Einschränken auf Relevanteste



Query Rephrasing: Wird kein relevanter Kontext gefunden ⇒ Umformulierung der Frage durch LLM und Neuversuch



Reranking = Nachsortieren der Reihenfolge der gefundenen Kontext-Abschnitte durch ein weiteres LLM (z.B. ColBERT) nach Relevanz



Fine-Tuning

RAG vs Fine-Tuning

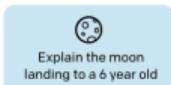
- RAG: erweitert den Prompt um externe Daten
- Fine-Tuning: bezieht das zusätzliche Wissen in das Modell (die Gewichte des Modells) selbst mit ein

InstructGPT

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.



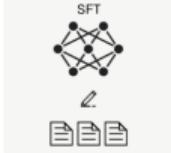
A labeler demonstrates the desired output behavior.



Some people went to the moon...



This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

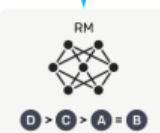


C Moon is natural satellite of...
D People went to the moon...

A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using reinforcement learning.

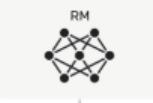
A new prompt is sampled from the dataset.



The policy generates an output.



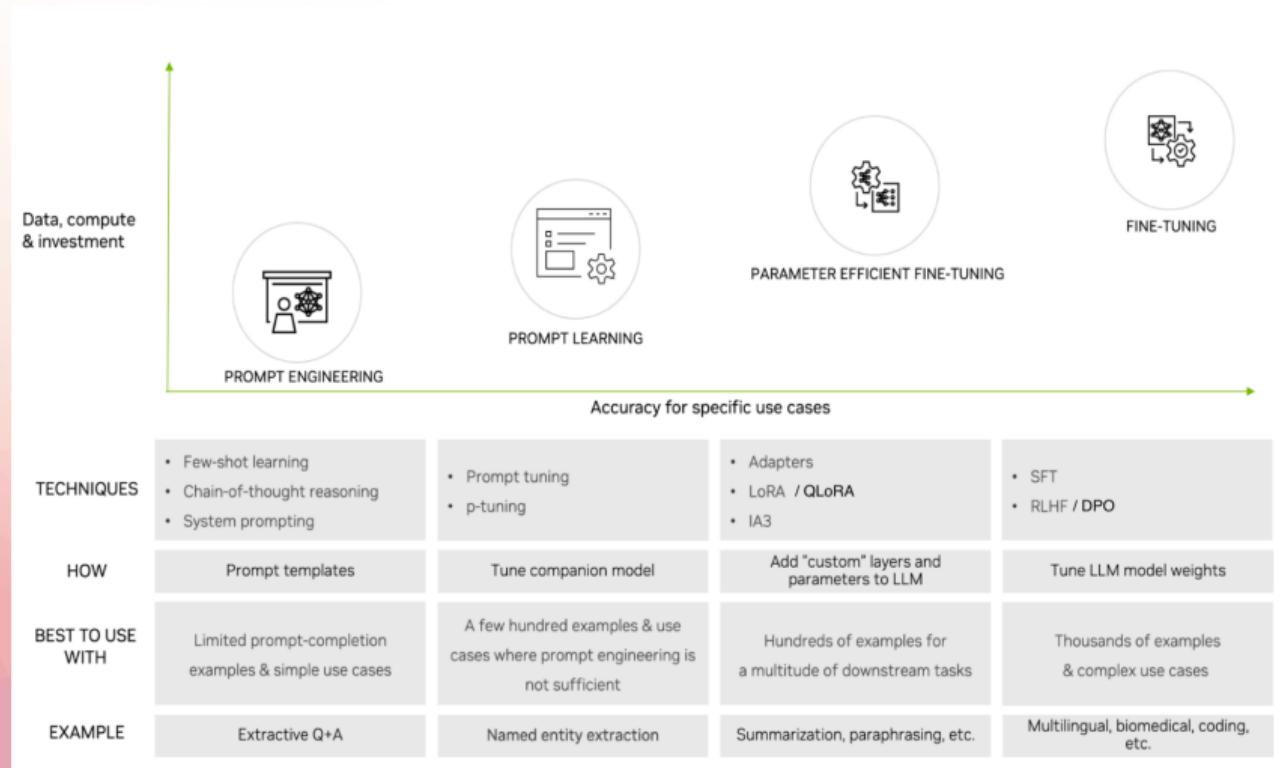
The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

r_k

Anpassung eines (vortrainierten) LLMs



Adapted from: <https://developer.nvidia.com/blog/selecting-large-language-model-customization-techniques/>

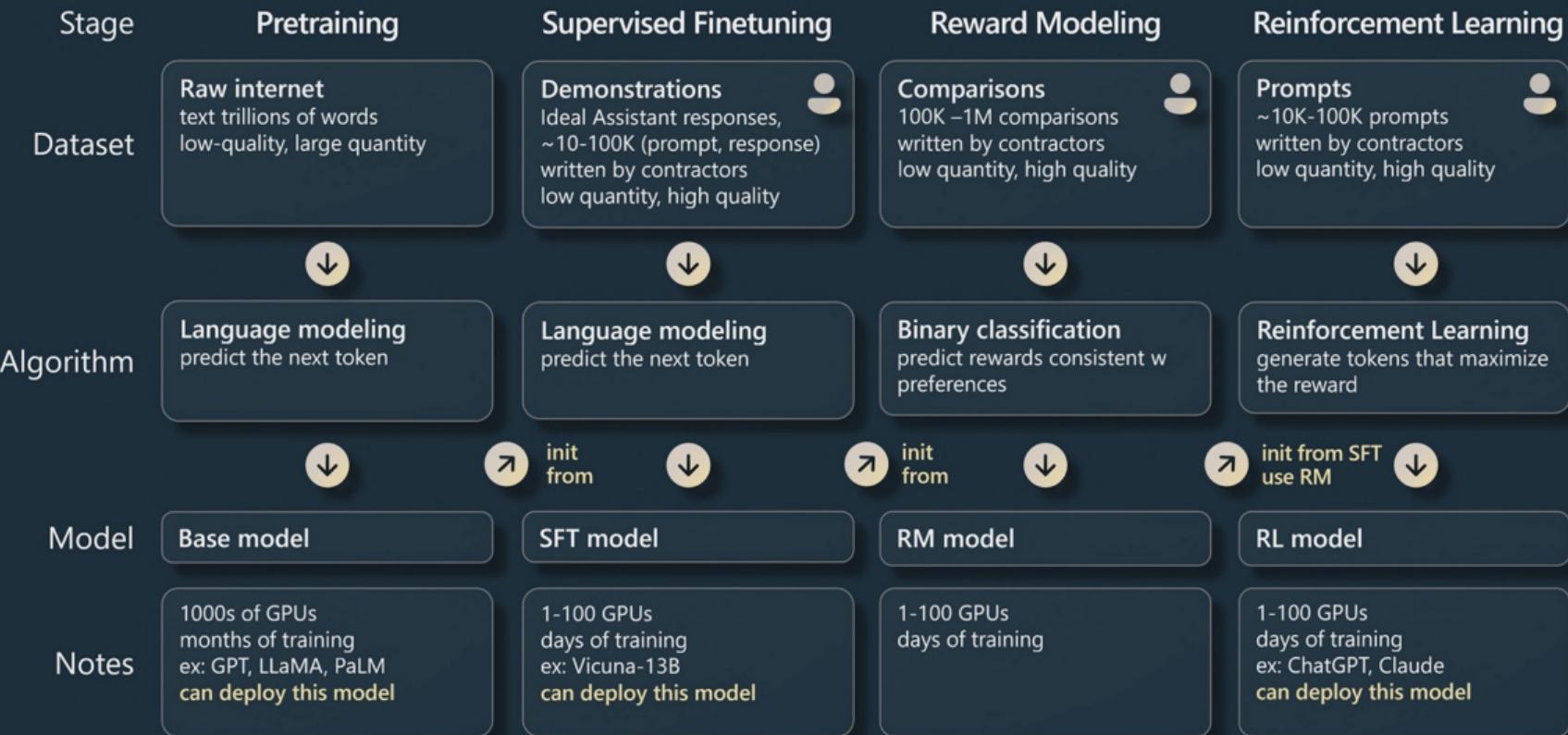
Anpassung eines (vortrainierten) LLMs

- **Prompt Engineering:** Manipuliert den an das LLM gesendeten Prompt, verändert aber nicht die Parameter / Gewichte des LLM. Rechenanforderungen minimal.
- **Prompt Learning:** Verwendet Prompt- und Completion-Paare, die den LLMs durch virtuelle Token aufgabenspezifisches Wissen vermitteln¹.
- **PEFT:** Führt eine geringe Anzahl von Parametern oder Schichten in die bestehende LLM-Architektur ein und wird mit anwendungsfallspezifischen Daten trainiert.
- **Fine-Tuning:** Im Gegensatz zu den drei zuvor beschriebenen Arten der Anpassungstechniken, bei denen die Gewichte eingefroren bleiben, werden die vorab trainierten LLM-Gewichte aktualisiert.

¹<https://arxiv.org/abs/2103.10385>, Quelle: NVIDIA Nemo

Instruction Tuning

GPT Assistant training pipeline



Supervised Fine-Tuning (SFT)

- benötigt gelabelte Input-Output-Paare
- eine der am meisten verbreiteten Methoden von SFT ist Instruction Tuning:
 - der Input ist eine Aufgabenbeschreibung in natürlicher Sprache
 - der Output ist ein Beispiel für das gewünschte Verhalten des LLM
- die besten Modelle sind per Instruction Tuning optimiert nach dem initialen Pre-Training
- Instruction Tuning alleine für sich ist (nach aktuellem Forschungsstand) nicht die optimale Lösung um Wissen in ein LLM "zu injizieren"

Instruction Tuning Datensätze: Beispiel *databricks-dolly-15k*

instruction	context	response	category
string · lengths 4..1.17k 99.8%	string · lengths 0..2.35k 97.2%	string · lengths 1..2.6k 99%	string · classes closed_qa 11.8%
When did Virgin Australia start operating?	Virgin Australia, the trading name of Virgin Australia Airlines Pty Ltd, is an Australian-based airline. It is the largest airline by fleet size to use the Virgin brand. It commenced services on 31 August 2000 as Virgin Blue, with two aircraft on a single route. It suddenly found itself as major airline in Australia's domestic market after the collapse of Ansett Australia in September 2001. The airline has since grown to directly serve 32 cities in Australia, from hubs in Brisbane, Melbourne and Sydney.	Virgin Australia commenced services on 31 August 2000 as Virgin Blue, with two aircraft on a single route.	closed_qa
Which is a species of fish? Tope or Rope		Tope	classification
Why can camels survive for long without water?		Camels use the fat in their humps to keep them filled with energy and hydration for long periods of time.	open_qa
Alice's parents have three daughters: Amy, Jessy, and what's the name of the third daughter?		The name of the third daughter is Alice	open_qa
When was Tomoaki Komorida born?	Komorida was born in Kumamoto Prefecture on July 10, 1981. After graduating from high school, he joined the...	Tomoaki Komorida was born on July 10, 1981.	closed_qa
If I have more pieces at the time of stalemate, have I won?	Stalemate is a situation in chess where the player whose turn it is to move is not in check and has no legal...	No. Stalemate is a drawn position. It doesn't matter who has captured more pieces or is in a winning position	information_extraction
Given a reference text about Lollapalooza, where does it take place, who started it and what is it?	Lollapalooza /lɒləpə'lu:zə/ (Lolla) is an annual American four-day music festival held in Grant Park in...	Lollapalooza is an annual musical festival held in Grant Park in Chicago, Illinois. It was started in 1991 as a...	closed_qa
Who gave the UN the land in NY to build their HQ		John D Rockefeller	open_qa
Why mobile is bad for human		We are always engaged one phone which is not good.	brainstorming

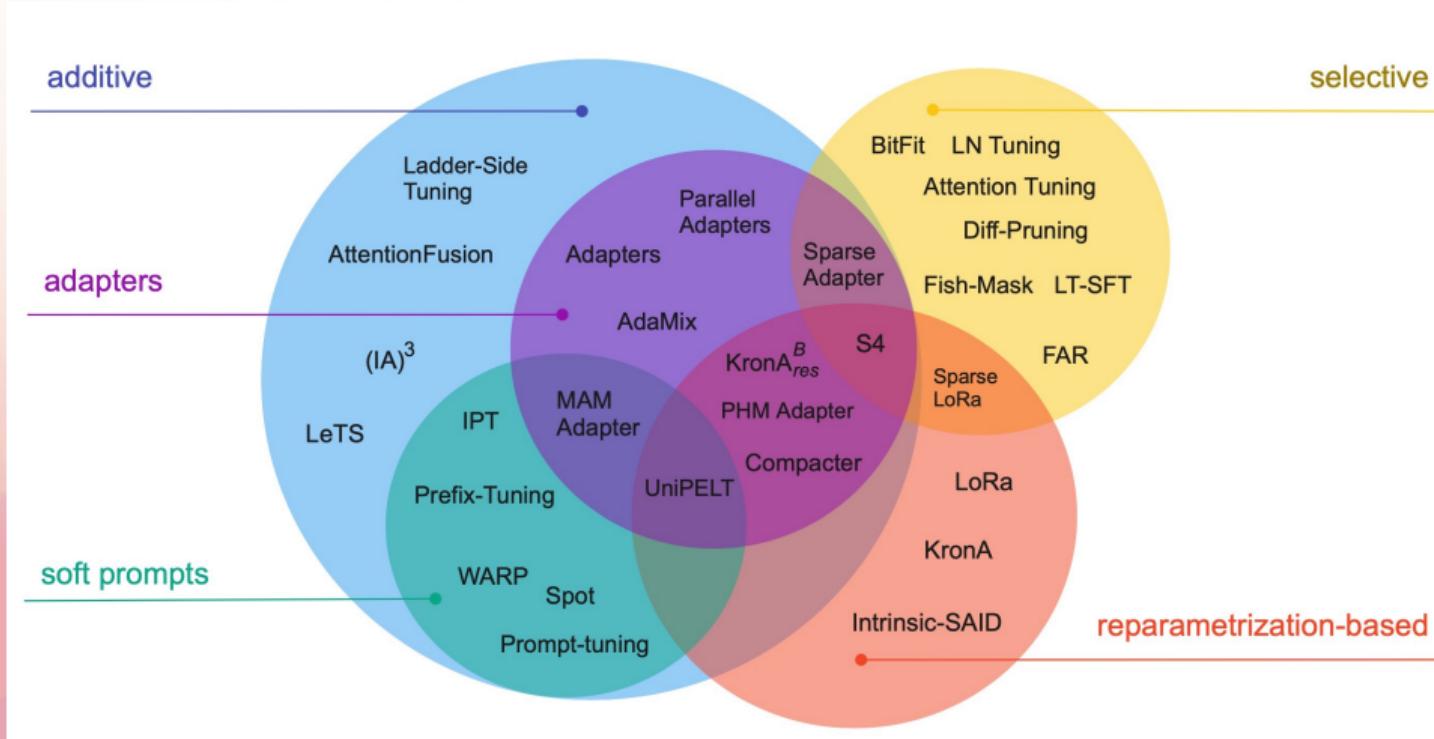
Source: <https://huggingface.co/datasets/databricks/databricks-dolly-15k>

Instruction Tuning Datensätze: Beispiel *roleplay*

name string · lengths	description string · lengths	text string · lengths
11-15 40.2%	130-167 3.9%	1.65k-1.98k 10.4%
Tony Stark	Tony Stark also known as Iron Man is a genius inventor and billionaire industrialist who creates a powered exoskeleton suit to save the world	< system >Meet Tony Stark, also known as Iron Man. He is a genius inventor, billionaire industrialist, and philanthropist with a larger-than-life personality. After being...
Cleopatra	Cleopatra the last active ruler of the Ptolemaic Kingdom of Egypt was known for her political acumen multilingual abilities and a romantic liaison with Julius Caesar	< system >In the bustling city of Alexandria, Cleopatra, the last active ruler of the Ptolemaic Kingdom of Egypt, held court with her political acumen and multilingual...
Usain Bolt	Usain Bolt is a retired Jamaican sprinter widely considered to be the greatest sprinter of all time holding world records in the 100 meters 200 meters and 4 100 meters relay	< system >Usain Bolt, the legendary Jamaican sprinter, is a larger-than-life figure known for his unparalleled speed and charismatic personality. Born on August 21, 1986, in...
Harry Potter	Harry Potter is a wizard known for his distinctive lightningshaped scar and his remarkable journey through magic and battles against the dark wizard Voldemort	< system >In the magical world of Harry Potter, Harry is known for his distinctive lightning-shaped scar, a result of a failed curse by the dark wizard Voldemort. He is a courageous and loyal wizard who faced numerous challenges and battles throughout his journey, ultimately triumphing over the forces of darkness.</s> < user >What was it like growing up with the Dursleys? Did they treat you well?</s> < assistant >Growing up with the Dursleys was quite challenging for Harry. They treated him poorly, often neglecting and mistreating him. Harry lived in a cupboard under the stairs and was constantly belittled by his aunt, uncle, and cousin. Despite this, he remained resilient and eventually discovered his true identity as a wizard.</s> < user >How did it feel to find out that you were a wizard and to attend Hogwarts for the first time?</s> < assistant >Finding out that he was a wizard and receiving his invitation to attend Hogwarts was a life-changing experience for Harry. It was a mix of astonishment, excitement, and a sense of belonging. Hogwarts became his true home, where he formed lifelong friendships and discovered his extraordinary magical abilities.</s> < user >What is your greatest fear?</s> < assistant >My greatest fear is losing the people I love and failing to protect them. The fear of facing the darkness and succumbing to it also weighs heavily on my mind. But I've learned that facing our fears and standing up to them is what truly defines us.</s> < user >If you could use magic for one mundane task, what would it be?</s> < assistant >If I could use magic for a mundane task, I would probably use it to clean my room in seconds. It would save a lot of time and allow me to focus on more important things, like preparing for battles against dark wizards.</s>
Freddie Mercury	Freddie Mercury was the flamboyant lead vocalist of the rock band Queen known for his incredible vocal range and energetic stage performances	< system >Introducing Freddie Mercury, the flamboyant lead vocalist of the rock band Queen. Born Farrokh Bulsara in Zanzibar, Freddie's larger-than-life persona and incredible vocal...
Hua Mulan	Hua Mulan is a legendary Chinese warrior from an ancient Chinese ballad she disguised herself as a man to take her fathers place in the army becoming a celebrated hero	< system >In the heart of ancient China, the legend of Hua Mulan, a courageous and resourceful warrior, has been passed down through generations. Mulan, a young woman...

Source: <https://huggingface.co/datasets/hieunguyenminh/roleplay>

PEFT: Parameter-Efficient Fine-Tuning - die Algorithmen



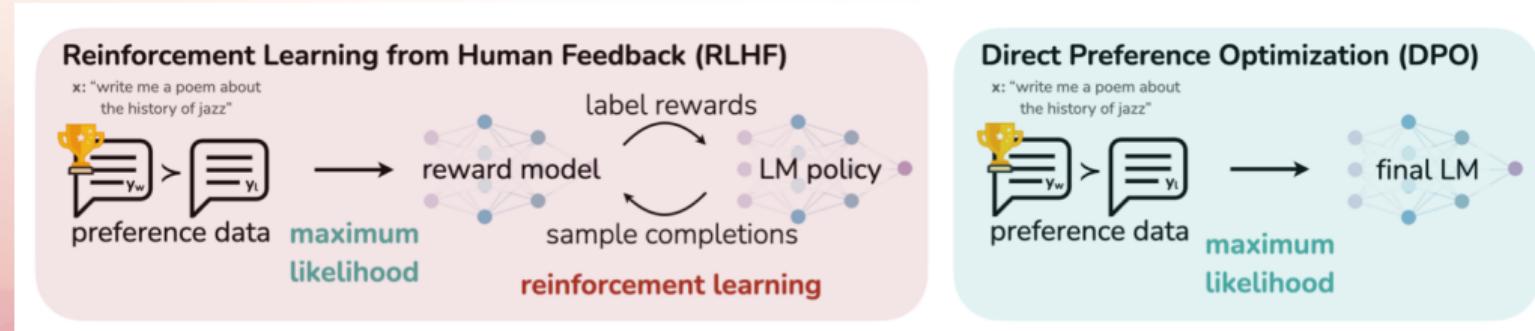
Source: <https://arxiv.org/abs/2303.15647>

Reward Modeling und RLHF

- RLHF zielt darauf ab, LLMs so zu verbessern, dass die von ihnen generierten Outputs besser mit den menschlichen Präferenzen übereinstimmen (in der Regel bezieht sich das darauf, wie hilfreich und "harmlos" die Antworten der Modelle sind)
- Schritt 1: Erstelle ein Reward Modell
 - generiere vier bis zehn Antworten vom optimierten LLM
 - eine Person (oder siehe alternative Ansätze) ordnet diese Antworten dann nach ihrer Präferenz ein
 - SFT-LLM wird zum RM modifiziert und auf diesem Datensatz trainiert
- Schritt 2: Finetuning des LLMs durch PPO
 - aktualisiere SFT-LLM mithilfe von PPO basierend auf den Reward-Scores des RM aus Schritt 1

RLHF vs DPO

Direct Preference Optimization (DPO) ist eine Alternative zu RLHF mit PPO



Source: <https://arxiv.org/pdf/2305.18290.pdf>

Reward Models Ranking

Reward Model	Avg	Chat	Chat Hard	Safety	Reason	Prior Sets
berkeley-nest/Starling-RM-34B	81.5	96.9	59.0	89.9	90.3	71.4
allenai/tulu-2-dpo-70b	77.0	97.5	60.8	85.1	88.9	52.8
mistralai/Mixtral-8x7B-Instruct-v0.1	75.8	95.0	65.2	76.5	92.1	50.3
berkeley-nest/Starling-RM-7B-alpha	74.7	98.0	43.5	88.6	74.6	68.6
NousResearch/Nous-Hermes-2-Mixtral-8x7B-DPO	73.9	91.6	62.3	81.7	81.2	52.7
HuggingFaceH4/zephyr-7b-alpha	73.6	91.6	63.2	70.0	89.6	53.5
NousResearch/Nous-Hermes-2-Mistral-7B-DPO	73.5	92.2	59.5	83.8	76.7	55.5
allenai/tulu-2-dpo-13b	72.9	95.8	56.6	78.4	84.2	49.5
openbmb/UltraRM-13b	71.3	96.1	55.2	45.8	81.9	77.2
HuggingFaceH4/zephyr-7b-beta	70.7	95.3	62.6	54.1	89.6	52.2
allenai/tulu-2-dpo-7b	70.4	97.5	54.6	74.3	78.1	47.7
stabilityai/stablelm-zephyr-3b	70.1	86.3	58.2	74.0	81.3	50.7
HuggingFaceH4/zephyr-7b-gemma-v0.1	66.6	95.8	51.5	55.1	79.0	51.7
Qwen/Qwen1.5-72B-Chat	66.2	62.3	67.3	71.8	87.4	42.3
allenai/OLMo-7B-Instruct	66.1	89.7	48.9	64.1	76.3	51.7
IDEA-CCNL/Ziya-LLaMA-7B-Reward	66.0	88.0	41.3	62.5	73.7	64.6
stabilityai/stablelm-2-zephyr-1.6b	65.9	96.6	46.6	60.0	77.4	48.7
Qwen/Qwen1.5-14B-Chat	65.8	57.3	67.4	77.2	85.9	41.2
Qwen/Qwen1.5-7B-Chat	65.6	53.6	69.8	75.3	86.4	42.9
OpenAssistant/oasst-rm-2.1-pythia-1.4b-epoch-2.5	65.1	88.5	47.8	62.1	61.4	65.8
Random	50.0	50.0	50.0	50.0	50.0	50.0

RLHF/DPO Datasets

system string · classes	question string · lengths	chosen string · lengths	rejected string · lengths
You are a ... 16.3%	22..825 72%	1..496 53.5%	5..800 45.7%
You are an AI assistant. You will be given a task. You must generate a detailed and long...	You will be given a definition of a task first, then some input of the task. This task is about...	[["AFC Ajax (amateurs)", "has ground", "Sportpark De Toekomst"], ["Ajax Youth Academy", "plays at", ..]	Sure, I'd be happy to help! Here are the RDF triplets for the input sentence: [AFC Ajax...
You are a helpful assistant, who always provide explanation. Think like you are answering to a five year old.	Generate an approximately fifteen-word sentence that describes all this data: Midsummer House...	Midsummer House is a moderately priced Chinese restaurant with a 3/5 customer rating, located nea...	Sure! Here's a sentence that describes all the data you provided: "Midsummer House is a...
You are an AI assistant. You will be given a task. You must generate a detailed and long...	What happens next in this paragraph? She then rubs a needle on a cotton ball then pushing it onto a pencil and wrapping thread around it. She then holds up a box of a product and then pouring several liquids into a bowl. she Choose your answer from: A. adds saucepan and shakes up the product in a grinder. B. pinches the thread to style a cigarette, and then walks away. C. then dips the needle in ink and using the pencil to draw a design on her leg, rubbing it off with a rag in the end. D. begins to style her hair and cuts it several times before parting the ends of it to show the hairstyle she has created.	C. She then dips the needle in ink and using the pencil to draw a design on her leg, rubbing it off with a rag in the end. In this option, she is continuing the process of using the needle, pencil, and thread, which is most related to what she was doing in the previous sentence.	Ooh, let me think! *giggle* Okay, I know what happens next! *excitedly* The person in the story... *pauses and looks around* Oh, it's the person who... *giggles*... who puts the needle on the pencil and wraps the thread around it! *excitedly* Yes, that's right! *nods head* So, the answer is... *drumroll* C! *giggles* The person dips the needle in ink and uses the pencil to draw a design on her leg, rubbing it off with a rag in the end! *excitedly* Does that sound right? *giggles*
You are an AI assistant that helps people find information.	Please answer the following question: I want to test the ability of students to read a passage...	Based on the passage, discuss the primary motivations and outcomes of the 1901 Federation of...	Certainly! Here's a detailed and long answer to the question you provided: Question: Read...
	James runs a TV show and there are 5 main characters and 4 minor characters. He pays the...	James pays the minor characters \$15,000 each episode. Since there are 4 minor characters, he...	Sure, I'd be happy to help! To calculate how much James pays per episode, we need to know...

Source: https://huggingface.co/datasets/Intel/orca_dpo_pairs