

Prototypische Analyse von KI-APIs für Large Language Models (Cloud vs. Onpremise)

Sandro Hartenstein

sandro.hartenstein@hwr-berlin.de

KI-Szenarien im Zeitalter von ChatGPT & Co

Fraunhofer IESE Kaiserlautern

21.11.2023



TAHAI

Agenda



- Motivation
- Large Language Models
- Prototypen
- Ergebnisse
- Fazit



Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Motivation

Motivation

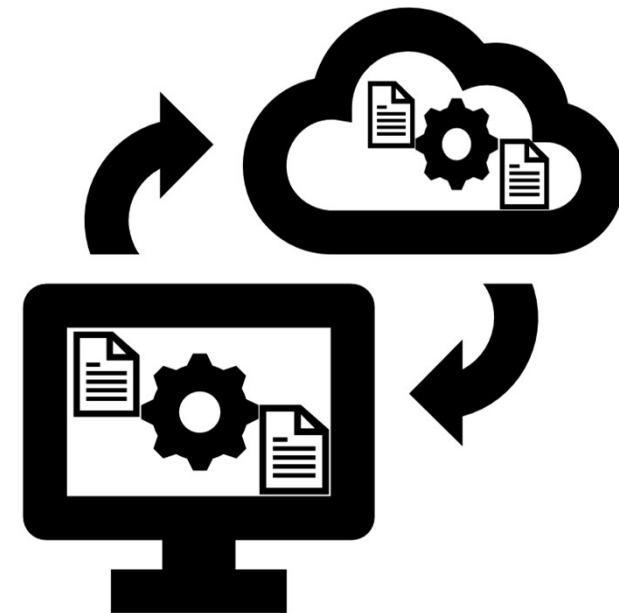


- Analyse von Mediationstranskriptionen
- Erwartungen an die KI-Analyse
 - Weniger Aufwand / Kosten
 - Objektivere Ergebnisse
 - Höherer Datenschutz



Motivation

- Ermitteln der Einstiegshürden für den Einsatz von LLM
- Neueste Entwicklungen prototypisch testen
- Möglichkeiten des lokalen Betriebes von LLMs untersuchen
- Potentiale von LLMs ermitteln

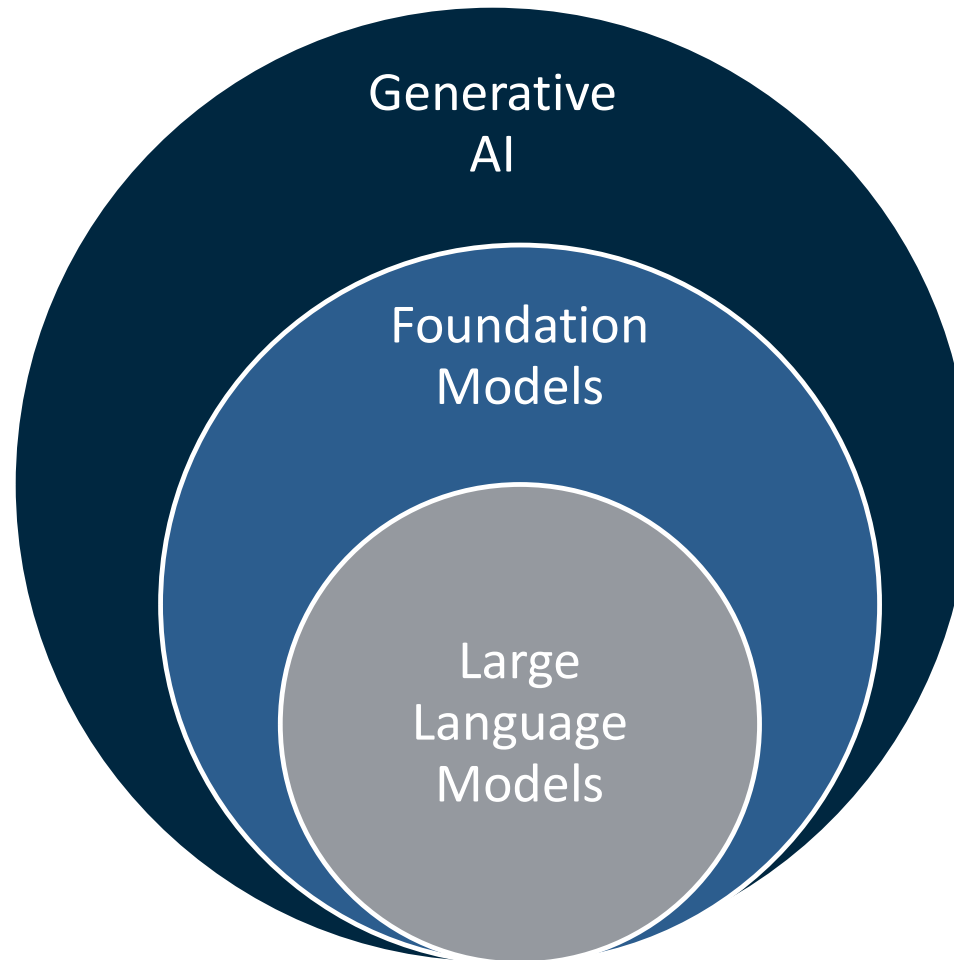




Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Large Language Models - LLMs

Large Language Models Einordnung



Large Language Models Komponenten



Daten

- Terrabytes an Daten
- Viele Parameter
 - GPT3.5 - 170B
 - LLAMA 2 - 70B



Architektur

- Neuronales Netz
 - GPT 2017
 - BERT 2018
- Schichten



Training

- Basistraining (nonSupervised)
- Finetuning (Supervised)
- Prompting





Large Language Models Funktionsweise (Transformerarchitektur)

Eingangsschicht:

Nimmt die Eingabedaten, wie z.B. Textsequenzen, entgegen.

Positional Encoding:

Fügt Informationen zur relativen oder absoluten Position der Eingaben hinzu, um die Sequenzinformationen zu erhalten.

Multi-Head-Attention:

Erlaubt dem Netzwerk, relevante Abhängigkeiten zwischen Wörtern in der Eingabe zu erkennen.

Feedforward-Schicht:

Führt eine Transformation auf den aus der Attention-Schicht kommenden Daten durch.

Normalisierungsschicht

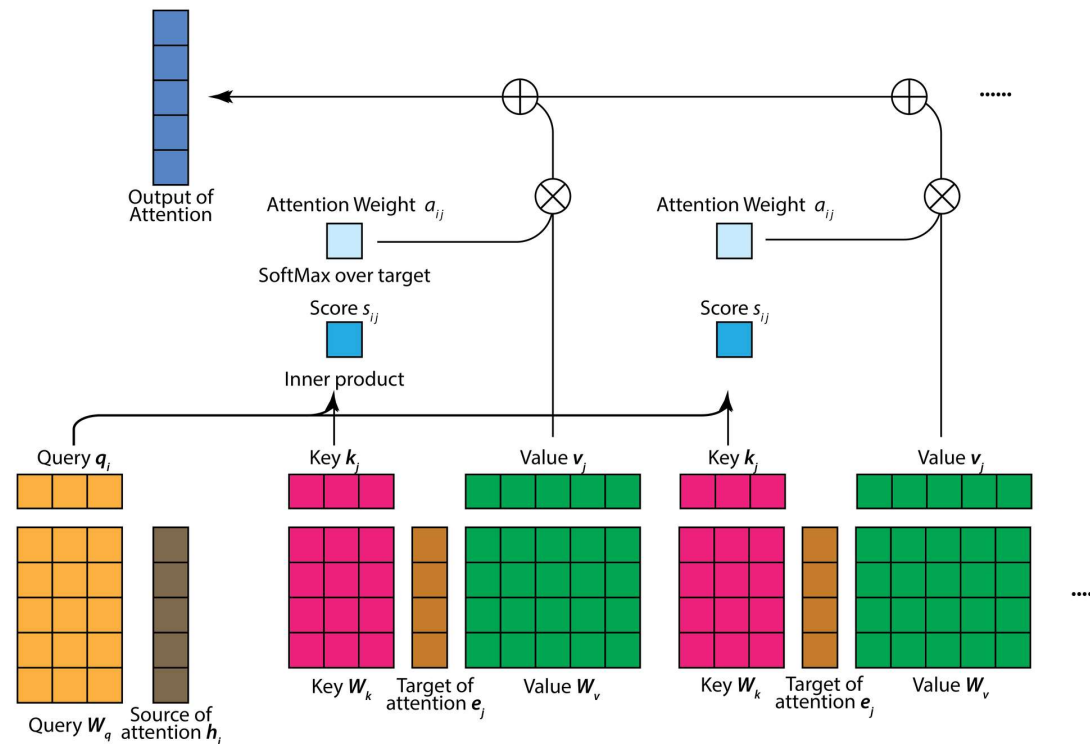
Gewährleisten eine stabilere Konvergenz während des Trainings.

Ausgangsschicht

Generiert die Ausgabe des Netzwerks, basierend auf den verarbeiteten Eingabedaten.

Large Language Models

Funktionsweise (Transformerarchitektur)

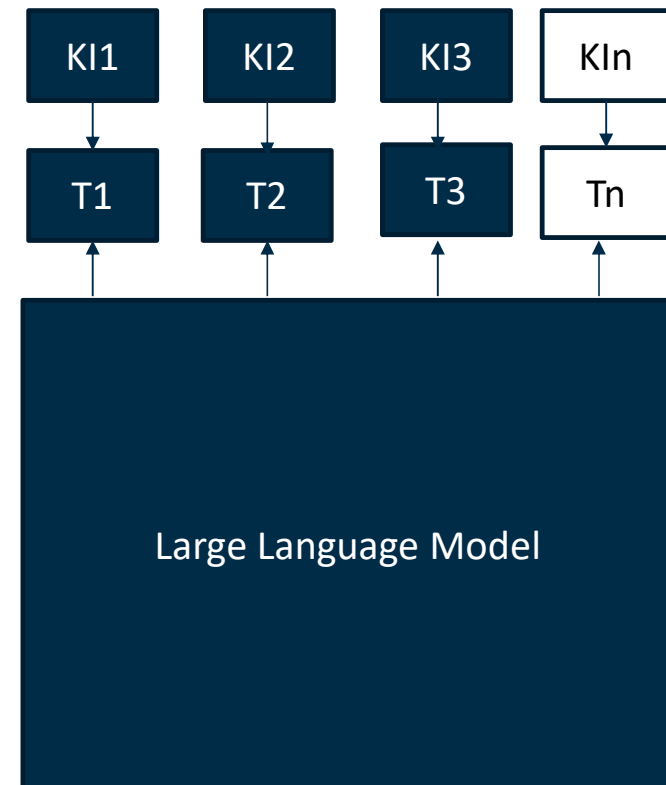


<https://scholar.harvard.edu/binxuw/classes/machine-learning-scratch/materials/transformers>



Large Language Models Anwendung

- Bisher spezielle KI-Modelle für jeweilige KI Aufgabe
 - Klassifikation
 - Sentiment
- Mit LLM nutzt man ein großes Modell für unterschiedlichste Aufgaben
 - Klassifikation
 - Sentiment
 - Assistants



Large Language Models Anwendung



Vorteile

- Bessere Ergebnisse ohne spezielles Training (ZeroShot)
- Allgemeine Anwendbarkeit
- Kontext
- Komplexere Aufgaben
- Prompting (Optimierung mit OneShot)

Nachteile

- Rechenressourcen (Betrieb, Training, Finetuning)
- Datenschutz
- Ethik
- Bias (je nach Trainingdaten)

Large Language Models

Anwendung

Typen



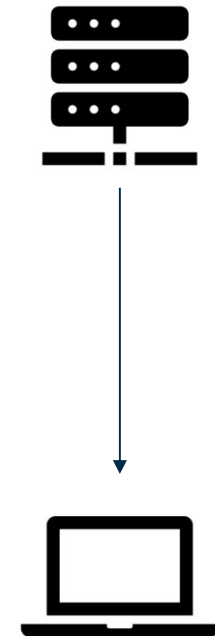
Base Models	Chat Models	Instuction Models
<ul style="list-style-type: none">• -> nur pretrained• (GPT3.5, BERT, LLAMA 2,)	<ul style="list-style-type: none">• -> finetuned für Konversation• (chatGPT, LLAMA 2 Chat)	<ul style="list-style-type: none">• -> finetuned für Generieren von Code / Aufgaben• (LLAMA Code , LLAMA 2 Instruct)

Large Language Models onpremis



LLMs auf PC Hardware?

- Rechenressourcen
 - Reduzierung der Parameter 7B, 13B, 70B statt. 170B
 - Reduzierung der Genauigkeit der Fließkommaberechnungen pro Parameter
 - FP16 bei LLAMA2 70B wären immer noch 140GB
 - Spezielle Quantisierungsmethoden (Q5 -> avg 5,5Bit per weight -> 48GB)
 - Nutzung von modernen CPUs (Multicore, Apple M1/2) und ggf. GPU
 - RAM intensiv (13B Q5_k_m-> 9GB, 70B Q5_k-m -> 48GB)
- Optimiertes Format GGUF





Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Prototypen

Prototypen Herausforderungen

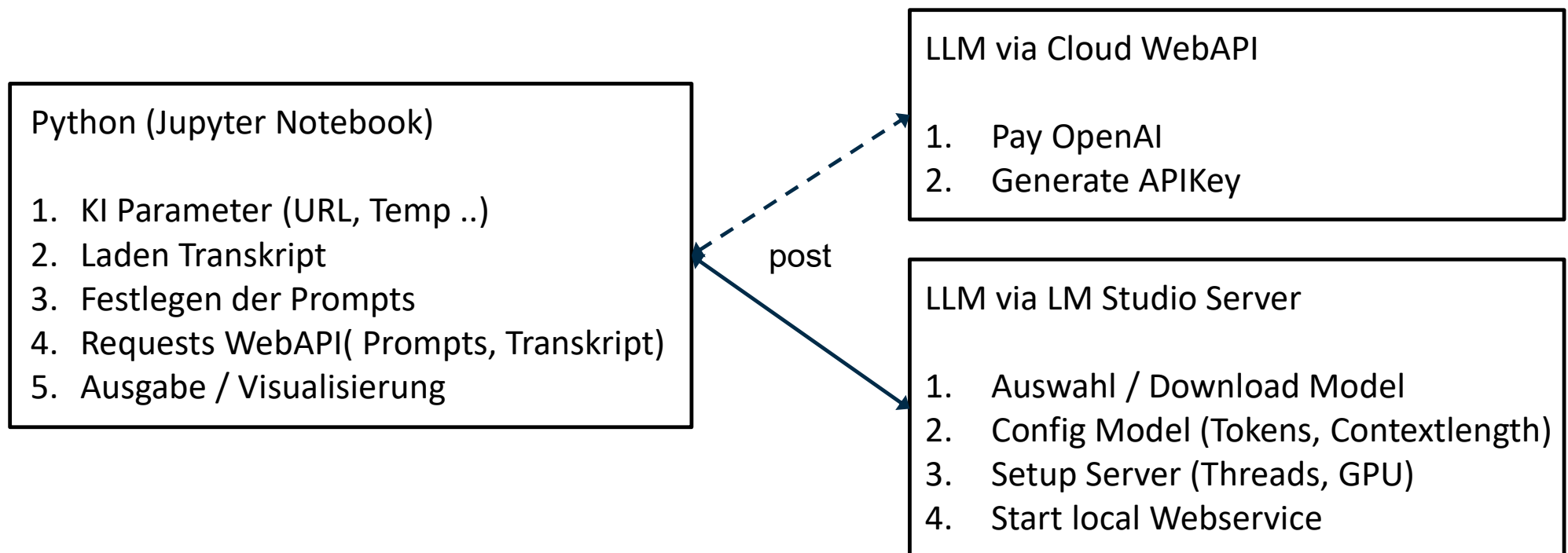


- Passendes LLM finden
 - Onprem <-> Cloud
 - Base <-> finetuned
 - Chat <-> Instruct
- Architektur & Code
 - WebAPIs <-> APIs
 - Python <-> Java
- Abfrage Setting (Prompt Design und Abfolge)
 - Sentiments über Zeitverlauf -> sequenzielle Analyse
 - Fragen zum Transkript -> kontextbasierte Analyse



Prototypen

Versuchsaufbau





Prototypen Implementierung - sequenzielle Analyse

```
def ask_ki(prompt):
    # Send the user prompt to ChatGPT using the correct endpoint for chat completions
    response = openai.ChatCompletion.create(
        model=model_name,
        messages=[{"role": "system", "content": "You are a Sentiment Analyzer. Always answer a number without text"},
                  {"role": "user", "content": prompt}],
        max_tokens=100,
        temperature = 0,
        top_k=3) # Adjust as needed
    return response
# Analyze sentiment for each line using the OpenAI API
for idx, line in enumerate(lines, start=1):
    match = re.match(r'^([\^:]+):', line)
    if match:
        current_speaker = match.group(1)
        prompt = f"{current_speaker}: {line}\n Rate the sentiment on a scale from -1 to 1: For example 0.25 "
        response = ask_ki(prompt)
        # Extract sentiment information from the response
        sentiment_value=extract_float(response['choices'][0]['message']['content'])
```



Prototypen

Implementierung - kontextbasierte Analyse

```
def ask_question_to_gpt(prompt):
    response = openai.ChatCompletion.create(
        model=model_name,
        messages=[
            {"role": "system", "content": "You are a helpful assistant and give short correct answers"},
            {"role": "user", "content": prompt}
        ],
        max_tokens=-1,
        temperature = temperature,
        top_k=3,
        context=context
    )
    return response['choices'][0]['message']['content'].strip()
context = conversation_text
questions = [ "Which mediation technique is used in the following conversation?" ]
temperature=0.0
for question in questions:
    answer = ask_question_to_gpt(f"Frage: {question}")
    print(f"Question: {question}\nAnswer: {answer}\n")
```

Prototypen Experimente (Beispieltranskriptionen)



Daten	Anwendung	NLTK- Polarity (Lagacy KI)	ChatGPT3.5	ChatGPT4	LLAMA2 13B	LLAMA2 70B
Gen. Mediation	Sentiment im Verlauf					
Gen. Mediation	Mediationstechnik / Offene Fragen	X				
Reale Mediation	Sentiment im Verlauf		X	X		
Reale Mediation	Mediationstechnik / Offene Fragen	X	X	X		



Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Ergebnisse



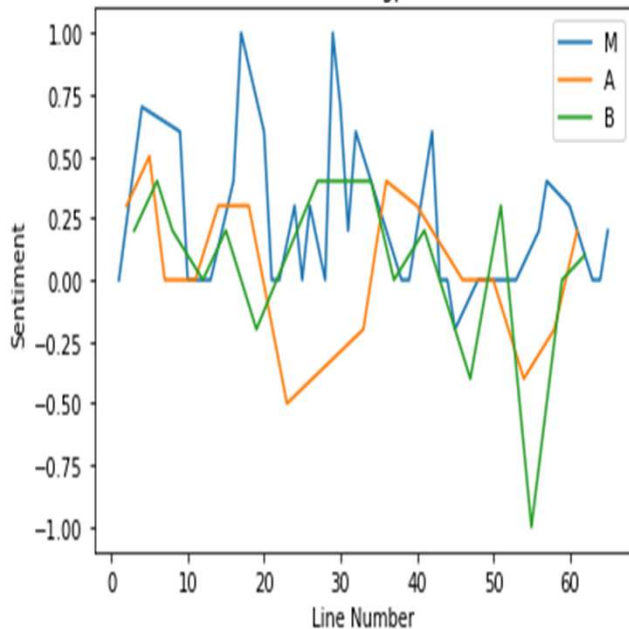
Ergebnisse

generierte Daten chatGPT 3.5 vs 4 vs LLAMA Chat 13B

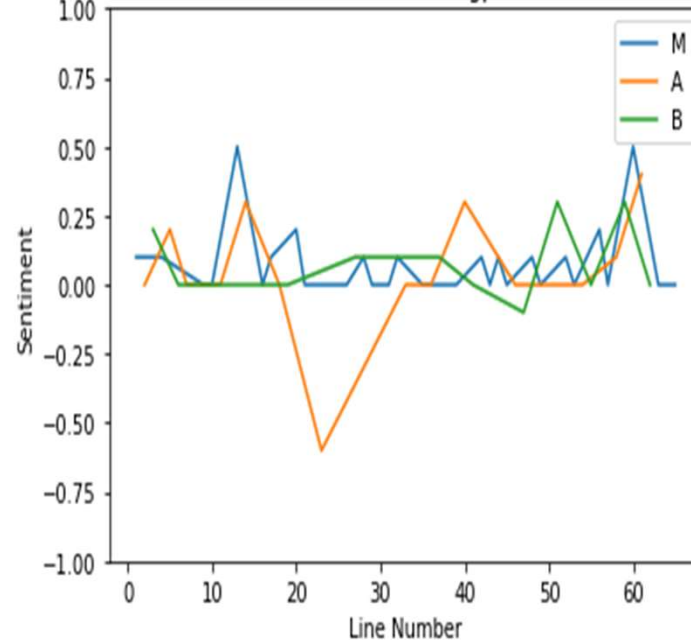
<https://chat.openai.com/share/9ced32e0-2565-4195-b41f-50b964c90994>

prompt = f"{current_speaker}: {line}\n Rate the sentiment on a scale from -1 to 1: For example 0.25 "

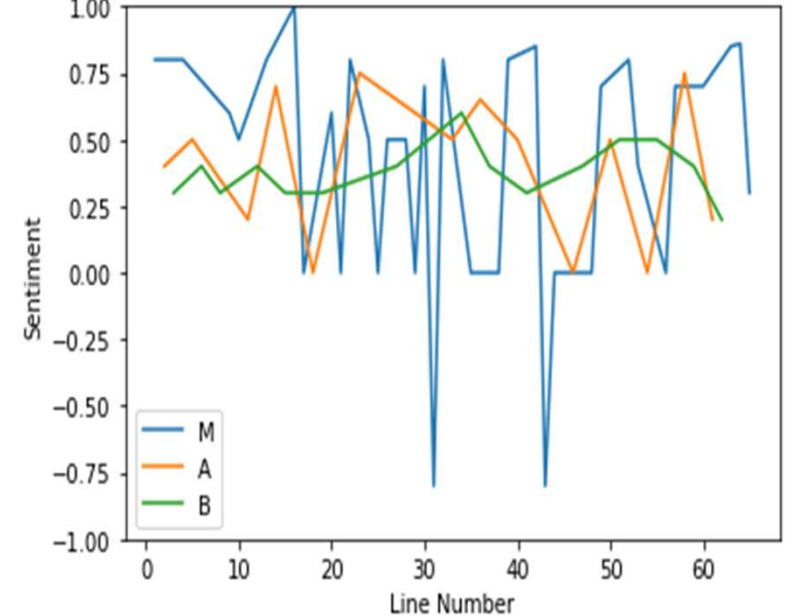
Sentiments Over Time - Model: gpt-3.5-turbo - File: med.txt



Sentiments Over Time - Model: gpt-4 - File: med.txt



Sentiments Over Time - Model: local-model - File: med.txt



Ergebnisse generierte Daten chatGPT 3.5 vs 4



identifizierte Mediationstechnik Model: gpt-3.5-turbo - File: med.txt :

Die Mediationstechnik, die in dieser Konversation verwendet wird, ist die "Reframing"-Technik. Der Mediator versucht, die Probleme neutral und objektiv zu formulieren, um die Emotionen etwas herauszunehmen und eine Lösungsorientierung zu fördern.

identifizierte Mediationstechnik Model: LLAMA

13B- File: med.txt :

The mediation technique used in this conversation is "Brainstorming".

identifizierte Mediationstechnik Model: gpt-4 - File: med.txt :

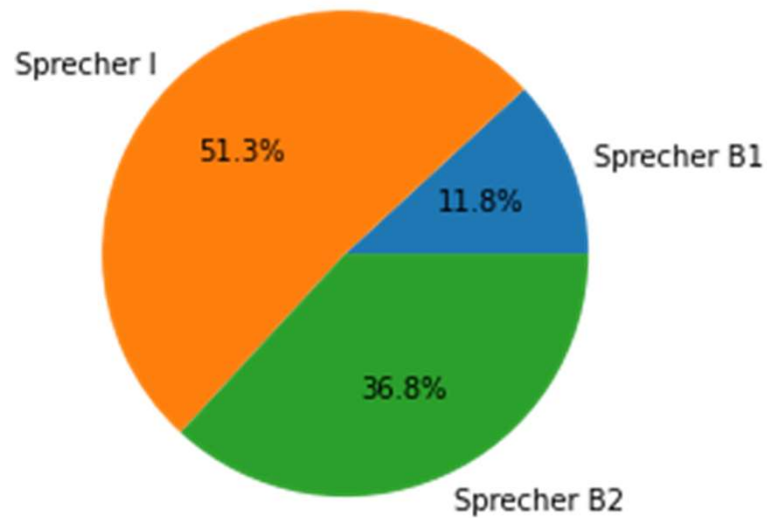
In dieser Konversation kommen verschiedene Mediationstechniken zum Einsatz: Es fängt mit der Strukturierung des Gespräches, geht über zur Caucus-Methode bei Einzelgesprächen, entwickelt dann Ideen durch Brainstorming und Schlussoptionen bei Implementierung eines Plans. Später wird die Reframing-Technik eingesetzt, um die Situation zu neutralisieren und schließlich wird die BATNA-Methode und die "Ja, aber"-Methode bei Verhandlung und Endvereinbarungen angewendet.



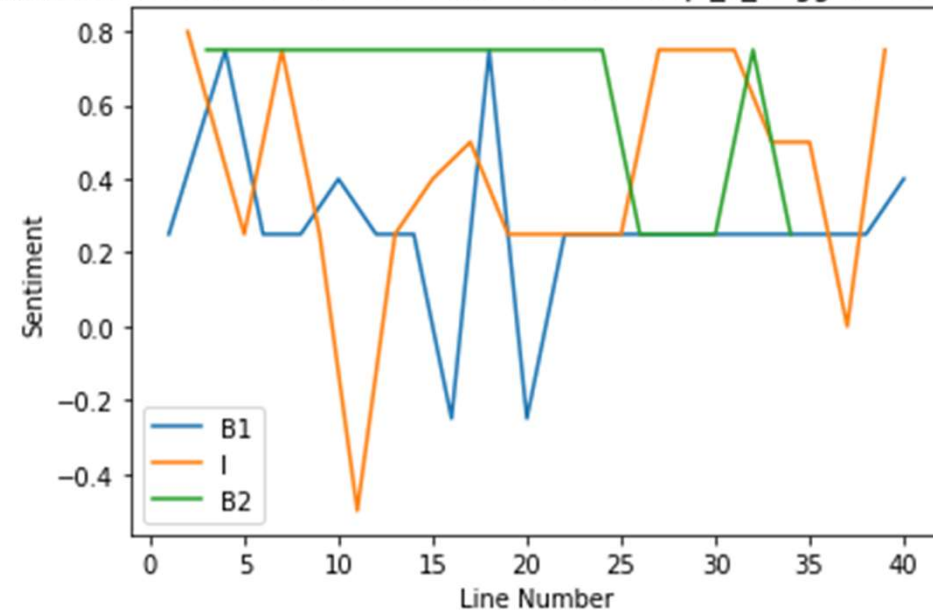
Ergebnisse EchtDaten

Model: llama 2 chat 70B q5_k_m gguf - File: realBsp.txt - Temp: 0.0 ZeroShot

Prozentualer Redeanteil nach Sprechern



Sentiments Over Time - Model: llama 2 chat 70B q5_k_m gguf - File: realBsp.txt





Ergebnisse

EchtDaten

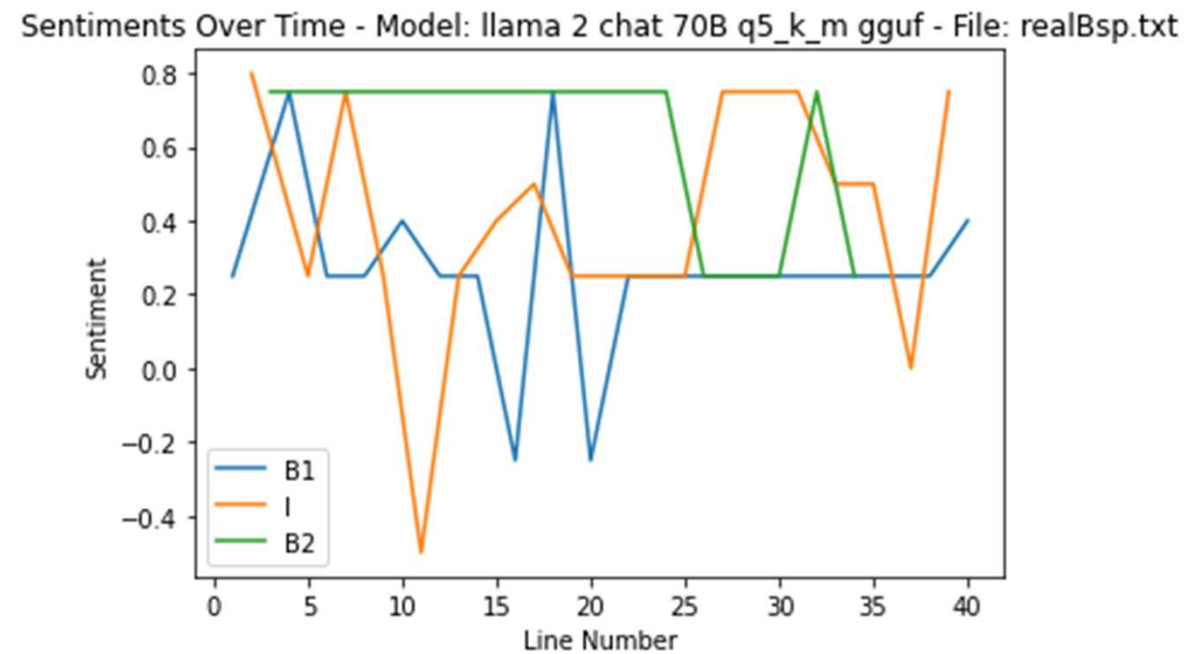
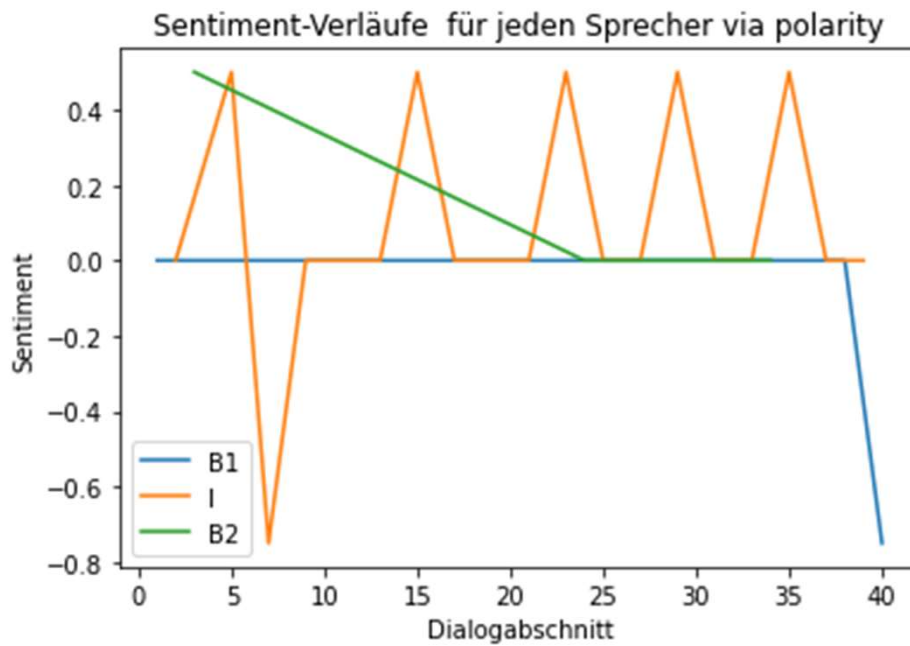
Model: llama 2 chat 70B q5_k_m gguf - File: realBsp.txt - Temp: 0.0

identifizierte Mediationstechnik:

The mediator is using a facilitative mediation approach, which focuses on helping the parties identify common goals and interests, communicate effectively, and find mutually beneficial solutions. The mediator is not taking a directive role, but rather asking open-ended questions to help the parties explore their perspectives and identify areas of common ground.

The mediator is also using active listening skills, such as paraphrasing and summarizing, to ensure that they understand the parties

Ergebnisse EchtDaten Polarity vs LLM





Ergebnisse

EchtDaten

Model: llama 2 chat 70B q5_k_m gguf - File: realBsp.txt - Temp: 0.0

Question: Which mediation technique is used in the following conversation?

Answer: Die Meditationsmethode, die in diesem Gespräch verwendet wurde, ist die Reflexionsfähigkeit.

Question: What were the main reasons for the conflict in the mediationtranscription? **Answer:** Die Hauptgründe für den Konflikt im Mediation

Question: How much of the conversation does each participant have?

Answer: In einem Gespräch haben die Teilnehmer je nach Bedarf und Thema unterschiedliche Mengen an Sprechzeit.

Question: Is the mediation technique appropriate, or is there room for improvement?

Answer: Mediation ist eine effektive Technik zur Konfliktlösung und kann dazu beitragen, Streitigkeiten zwischen Parteien angemessen zu lösen. Es gibt jedoch immer noch Raum für Verbesserungen, um die Effektivität der Mediation weiter zu steigern.

Ergebnisse Bewertung



Daten	Anwendung	Polarity (Legacy KI)	ChatGPT3.5	ChatGPT4	LLAMA2 13B	LLAMA2 70B
Gen. Mediation	Sentiment im Verlauf	Sehr schwach	Gut (Schwankung)	Sehr gut	Schwach (sehr starke Schwankung)	x
Gen. Mediation	Mediationstechnik / Offene Fragen	X	gut	gut	Schwach „Brainstorming“ / gut	x
Reale Mediation	Sentiment im Verlauf	Sehr schwach	X	X	X	Gut
Reale Mediation	Mediationstechnik / Offene Fragen	X	X	X	X	Gut „Reflexionsfähigkeit“ / schwach



Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Fazit



Fazit

Lessons-Learned

- LLMs sind sehr leistungsfähig in Textanalysen von Transkripten
- Lokale LLMs recht einfach zu verwenden (LM Studio)
- Sequenzielle Analysen mit LLMs sind sehr zeitintensiv
- Kontextbasierte Analysen bieten Zugang für Domänenexperten
- Prompting bietet viel Potential

Fazit

Ausblick



Weiterentwicklung des Prototypen

- Useability (Webfrontend)
- LLM Auswahl optimieren
- Domainspezifischere Evaluation
 - Manuelle Analysen vs Automatisierte Analysen
- ggf. Foundation Modell für Audioanalysen
 - Identifikation von Zwischentönen und besondere Audiomerkmale

Vielen Dank!
Fragen?

Sandro Hartenstein
sandro.hartenstein@hwr-berlin.de

KI-Szenarien im Zeitalter von ChatGPT & Co
Fraunhofer IESE Kaiserlautern
21.11.2023



TAHAI

Transformer Architektur



<https://www.wolfram.com/language/12/neural-network-framework/use-transformer-neural-nets.html.de>

<https://arxiv.org/abs/1706.03762> -- Attention is all you need - Google 2017

https://scholar.harvard.edu/sites/scholar.harvard.edu/files/binxuw/files/mlfs_tutorial_nlp_transformer_ssl_updated.pdf