

Christian Lang

RETRIEVAL AUGMENTED GENERATION (RAG) – EXTERNE WISSENSBASEN FÜR LARGE LANGUAGE MODELS AM BEISPIEL DEUTSCHER GRAMMATIK

Webinar für DHd-AG Angewandte Generative KI in den Digitalen Geisteswissenschaften

WORÜBER REDEN WIR?

- LLM für QA
 - Bedingungen für produktiven Einsatz
- Retrieval Augmented Generation (RAG)
- Use Case *grammis*
 - Evaluationsstudie
- Weitere Schritte und Fazit

LLM FÜR QA

- **Potenziale** von Large Language Models (LLM) sind nahezu unendlich, eine naheliegende Anwendung: automatische **QA-Systeme** („Chatbots“)
- Unser **Use Case**: QA-System als Erweiterung des grammatischen Informationssystems **grammis** (<https://grammis.ids-mannheim.de>).
- Ein solcher (möglicher) produktiver Einsatz eines LLM bringt eine Reihe von **Bedingungen** mit sich...



Frage der Woche



Wir haben kommen müssen oder kommen gemusst, ich habe früh lesen können oder lesen gekonnt? – Perfekt von Modalverben

Inhaltliche Korrektheit/Konsistenz

- **ABER:** Auf fachliche Fragen antworten LLM oft **vage** und **widersprüchlich**, bisweilen **halluzinieren** die Modelle (vgl. <https://dhd-blog.org/?p=19645>).



Was ist eine Abtönungspartikel?



Eine Abtönungspartikel (auch als article of diminution bezeichnet) ist ein Wortartikel, der in einigen germanischen Sprachen wie Deutsch, Niederländisch, Schwedisch und Dänisch verwendet wird. Er dient dazu, ein Substantiv abzutönen, das heißt, er reduziert die Intensität oder den Schweregrad des Substantivs.

In Deutsch wird der Abtönungspartikel meist als "lein" oder "chen" verwendet. Zum Beispiel:

- * Hund → Hündlein
- * Haus → Häuschen
- * Mann → Männlein

Der Abtönungspartikel kann verwendet werden, um ein Substantiv zu verkleinern, um eine süße oder naive Konnotation hinzuzufügen oder um ein Gefühl von Intimität oder Nähe zu erzeugen. Er kann auch verwendet werden, um ein Substantiv in eine Form zu bringen, die für Kinder oder Tiere geeignet ist.

Es ist jedoch zu beachten, dass der Gebrauch von Abtönungspartikeln in modernen Deutsch sprechenden Regionen nicht sehr häufig ist und oft als altertümlich oder übertrieben empfunden wird.

Inhaltliche Korrektheit/Konsistenz

- **ABER:** Auf fachliche Fragen antworten LLM oft **vage** und **widersprüchlich**, bisweilen **halluzinieren** die Modelle (vgl. <https://dhd-blog.org/?p=19645>).



Eine Abtönungspartikel ist ein Wort oder eine Wortgruppe in einer Aussage, die die Aussage selbst nicht ändert, aber den Ton oder die Nuance der Aussage beeinflusst. Abtönungspartikel können die Aussage mildern, betonen, relativieren oder ironisch färben.

Beispiele für Abtönungspartikel sind:

- * "eigentlich" (mildert die Aussage ab)
- * "natürlich" (betont die Selbstverständlichkeit)
- * "irgendwie" (relativiert die Aussage)
- * "sicher" (betont die Sicherheit)
- * "ja" (kann die Aussage bestätigen oder ironisch färben)
- * "doch" (kann die Aussage betonen oder widersprechen)

Abtönungspartikel können in verschiedenen Kontexten verwendet werden und sind wichtig für die Kommunikation, da sie helfen, die Intention und den Ton einer Aussage zu vermitteln.

Nachvollziehbarkeit

- Woher stammt die einer Antwort zugrundeliegende Information?

ABER: Quellenlange intransparent



Woher stammen diese Informationen?



Die Informationen, die ich Ihnen über Abtönungspartikel bereitgestellt habe, stammen aus meiner Ausbildung und meinem Training im Bereich der Linguistik und der deutschen Grammatik.

Konkret basieren meine Antworten auf folgenden Quellen:

- * Deutsche Grammatikbücher und -lehrbücher, wie z.B. "Duden - Die Grammatik" oder "Grammatik der deutschen Sprache" von Helmut Glück
- * Fachartikel und Studien aus der Linguistik, wie z.B. "Abtönungspartikel im Deutschen" von der Universität Leipzig
- * Online-Ressourcen und -Wörterbücher, wie z.B. Wikipedia oder das DWDS-Wörterbuch

Ich habe diese Informationen verarbeitet und in einer verständlichen und zusammenfassenden Form präsentiert, um Ihnen zu helfen, das Konzept der Abtönungspartikel zu verstehen.

Aktualität

Neue Erkenntnisse sollten repräsentiert sein.

ABER: Was nicht in den Trainingsdaten repräsentiert ist, steht für Antworten nicht zur Verfügung



Wer ist deutscher Bundeskanzler?



Das ist eine gute Frage!

Aktuell (Stand 2023) ist Olaf Scholz der deutsche Bundeskanzler. Er ist seit dem 8. Dezember 2021 im Amt und führt eine Ampelkoalition aus SPD, FDP und Bündnis 90/Die Grünen.

Bitte beachten Sie, dass die politische Lage sich ändern kann, und es ist immer ratsam, sich an vertrauenswürdige Quellen zu wenden, um aktuelle Informationen zu erhalten.

BEDINGUNGEN UND HERAUSFORDERUNGEN

Aktualität

Neue Erkenntnisse sollten repräsentiert sein.

ABER: Was nicht in den Trainingsdaten repräsentiert ist, steht für Antworten nicht zur Verfügung



Wer ist deutscher Meister?



Eine Frage, die das Herz jedes Fußballfans höher schlagen lässt!

Aktuell ist der FC Bayern München deutscher Meister. Sie haben in der Saison 2021/22 die Bundesliga gewonnen und sind damit zum 32. Mal deutscher Meister geworden.

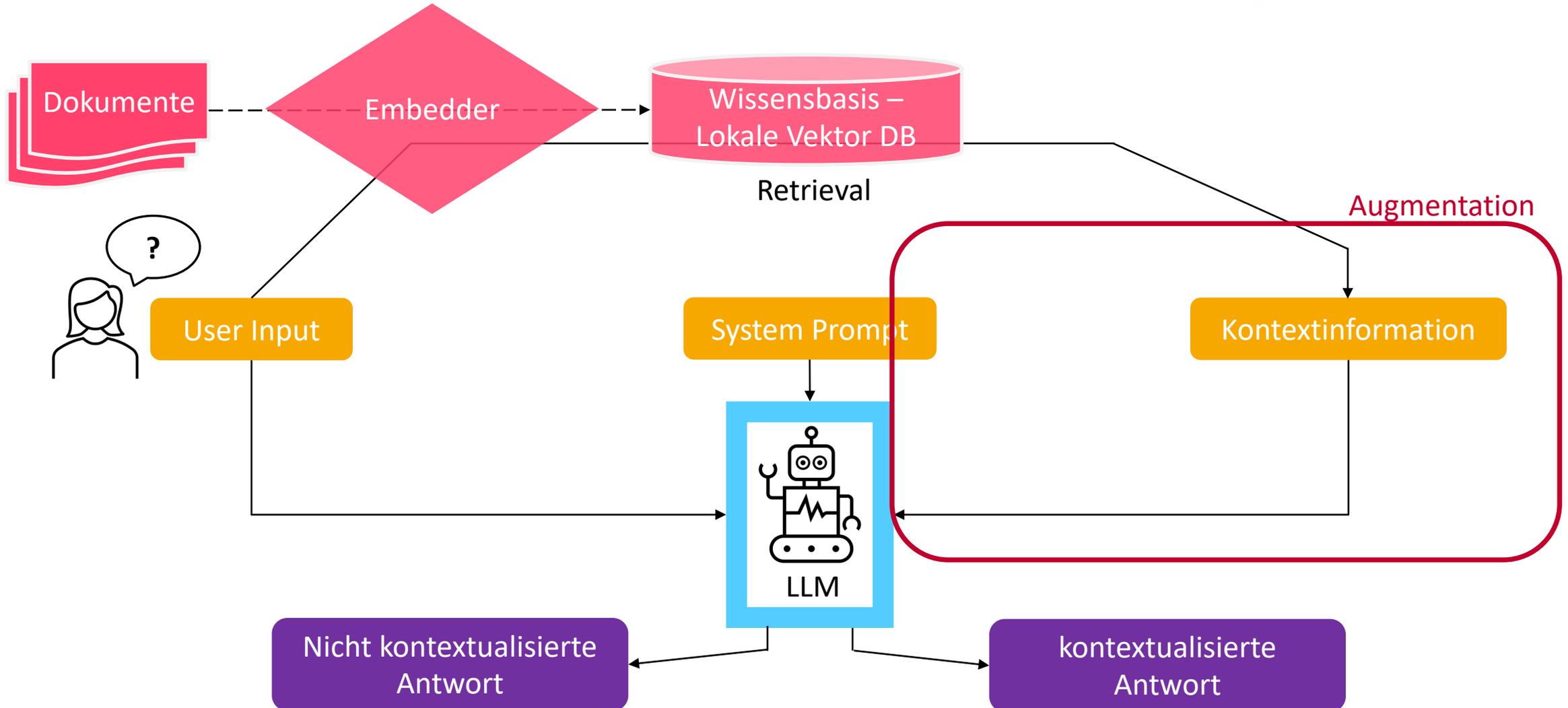
Allerdings muss ich darauf hinweisen, dass die Frage "Wer ist deutscher Meister?" je nach Saison und Jahr eine unterschiedliche Antwort haben kann. Wenn du also wissen möchtest, wer in einer bestimmten Saison deutscher Meister war, stehe ich gerne zur Verfügung, um dir die Antwort zu liefern!

RETRIEVAL AUGMENTED GENERATION (RAG)

- **mögl. Lösung** besteht darin, dem LLM „kontrollierten“ Zugriff auf Spezialwissen zu ermöglichen.
- Retrieval Augmented Generation (RAG) kombiniert die **generative** Fähigkeit eines LLM mit einem **Retrievalsystem**.
- Grundlegendes Prinzip: Das LLM erhält zu jeder Frage relevante Informationen für deren Beantwortung.
- **Open Book** Paradigma (vs. Closed book)



ARCHITEKTUR RAG

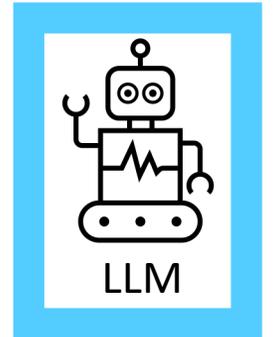


- RAG macht „**externes**“ **Wissen für LLM zugänglich**. Das vortrainierte LLM selbst – im Sinne seiner Gewichte – bleibt unverändert.
- Finetuning beschreibt den Prozess, in dem ein vortrainiertes LLM im Sinne von Gewichtsanpassungen **weiter trainiert** wird.
- Finetuning benötigt **immense Ressourcen** (Zeit, Rechenpower etc.)
- Für unseren Task (Akkuratheit, Vermeidung von Halluzinationen) scheint RAG eine geeignete Methode.
- Finetuning für Erlernen neuer Tasks, spezielle „Sprechweisen“

- **Lokaler Betrieb: Systemvoraussetzungen** für Betrieb sind **enorm**. Es gibt **Cloud-Services**, jedoch: Nicht in jeder Situation ist es gewollt, dass Daten das eigene Ökosystem verlassen und an **externe Server** geschickt werden.
- **Quantisierung** und kleinere **Modellgrößen** als Möglichkeit, LLM auf handelsüblicher Hardware (und damit lokal) zu betreiben
- Grundidee von Quantisierung: **Reduzierung der Modellgröße durch Reduzierung der Anzahl Bits für Modellgewichte**, indem die Präzision der zugrundeliegenden Datentypen reduziert wird, bspw. 16-bit floating point (16 FP) zu 4-bit Integer (INT4).

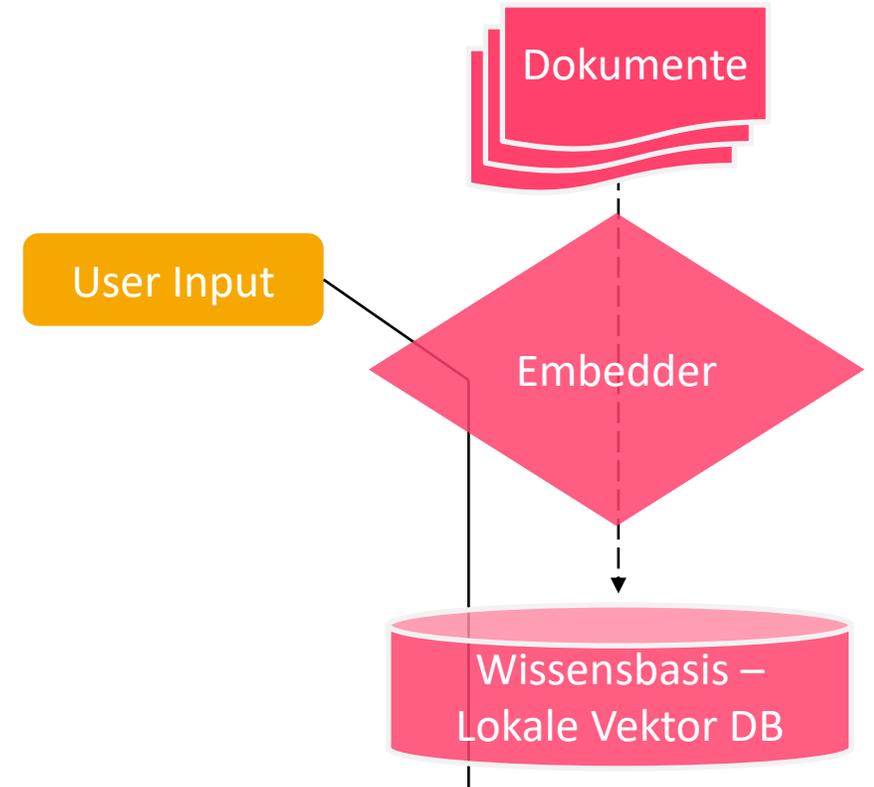
USE CASE GRAMMIS

- Wir verwenden Llama 2 mit 13B Parametern und Chat-Finetuning in 4-bit Quantisierung (mittlerweile ist Llama 3 verfügbar).
- Llama ist ein LLM von Meta und ist „**frei verfügbar für Forschung und kommerziellen Gebrauch.**“ <https://ai.meta.com/llama/>
- Es ist verfügbar in drei **Größen**: mit 7B, 13B und 70B Modellparametern (mit und ohne Chat-Finetuning)...
- Auf 🥰Huggingface (<https://huggingface.co>) finden sich unzählige Varianten (Tipp: User <https://huggingface.co/TheBloke>)
- Zum Zeitpunkt unserer Evaluation war Llama 2 vergleichbar performant wie GPT 3.5



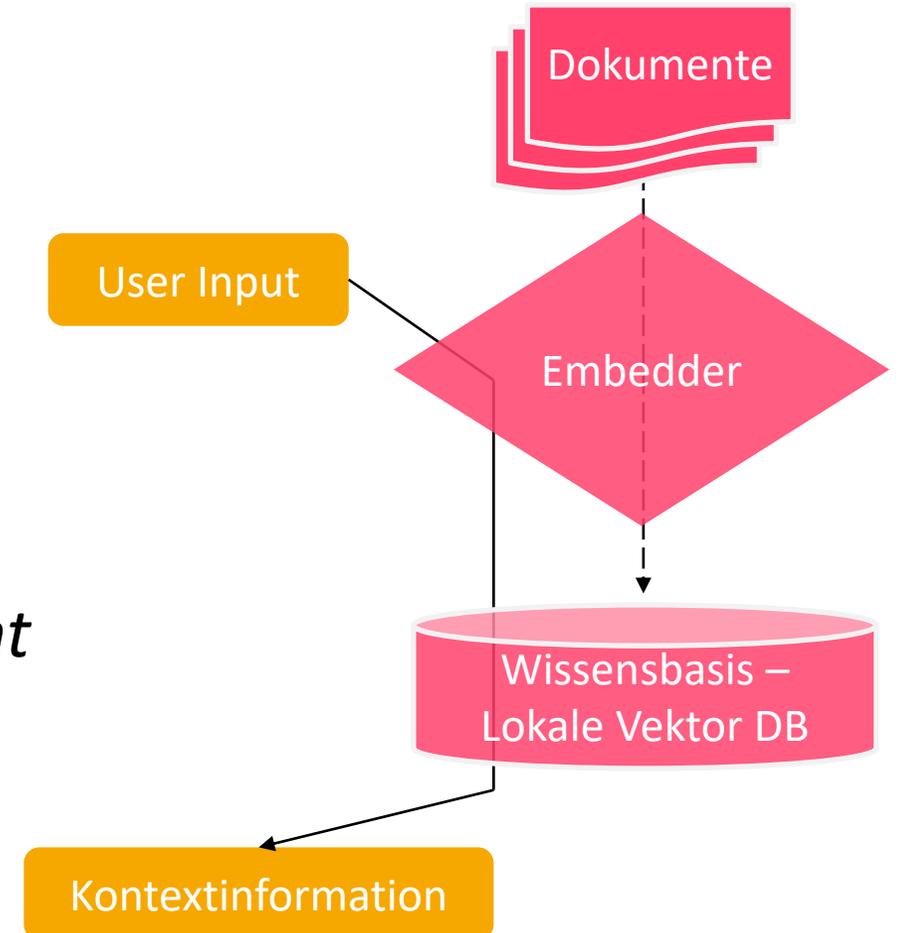
USE CASE GRAMMIS

- **Pythonbasierte Toolchain:**
<https://github.com/PromptEngineer/localGPT>
(verbindet u.a.: LangChain, InstructorEmbeddings, ChromaDB, bitsandbytes, llama.cpp etc.)
- **Dokumente für Wissensbasis:** Textuelle Inhalte von *grammis* (ca. 1 Mio. Token) und Fragen + Antworten, die an einen kommerziellen Sprachberatungsservice gesendet wurden (ca. 6 Mio. Token)
- Embedding der Dokumente und des User-Inputs mittels *instructor-xl* (<https://huggingface.co/hkunlp/instructor-xl>)



USE CASE GRAMMIS

- Ermittlung zum Input passender Dokumente in der Wissensbasis durch Kosinusähnlichkeit.
- Möglichkeit, die relevanten Dokumente auszugeben: Transparenz und Kontrollmöglichkeit der Retrievalqualität
- System Prompt: „*Use the following context to answer the question. If you do not know the answer, say that you do not know. Do not invent answers. Translate your answer into German.*“
- **Augmentation:** System Prompt + **Kontextinformation** + Input (Frage)



DEMO

```
(localGPT_llama) C:\python_skripte\localGPT_llama>python run_localGPT.py
```

„Abtönungspartikel sind Partikel, die die Erwartungen und Einstellungen des Sprechers und des Hörers bei der Kommunikation widerspiegeln. Es gibt verschiedene Typen von Abtönungspartikeln, die formal unterschiedlich sind. Abtönungspartikeln operieren auf Einstellungen und können nicht erfragt, nicht negiert und nicht als Kopf einer Phrase sein. Sie sind unflektierbar und können nur nach bestimmten Regeln kombiniert werden.“

- Diese von Llama 2 13B 4-bit erstellte Antwort ist deutlich besser als die eingangs gezeigte von Llama 3 70B ohne Quantisierung.

AUSBLICK: EVALUATIONSTUDIE

- Systematische Evaluierung der Antworten der lokalen Implementierung RAG-Framework (*4bit_RAG*) gegen zwei Baselines (*4bit* und Llama 13B Chat ohne Quantisierung: *no_quant*).
- Ausgewertet wurden 24 Evaluationsfragen basierend auf den authentischen Daten des Sprachberatungsservices (diese waren nicht Teil der Wissensbasis).
- Differenzierung in zwei verschiedene Fragearten:
 - **Definitorische Fragen** („*Was ist eine Abtönungspartikel?*“)
 - **Gebrauchsfragen** („*Schreibt man ‚Onlineshop‘ mit oder ohne Bindestrich?*“)

AUSBLICK: EVALUATIONSSTUDIE

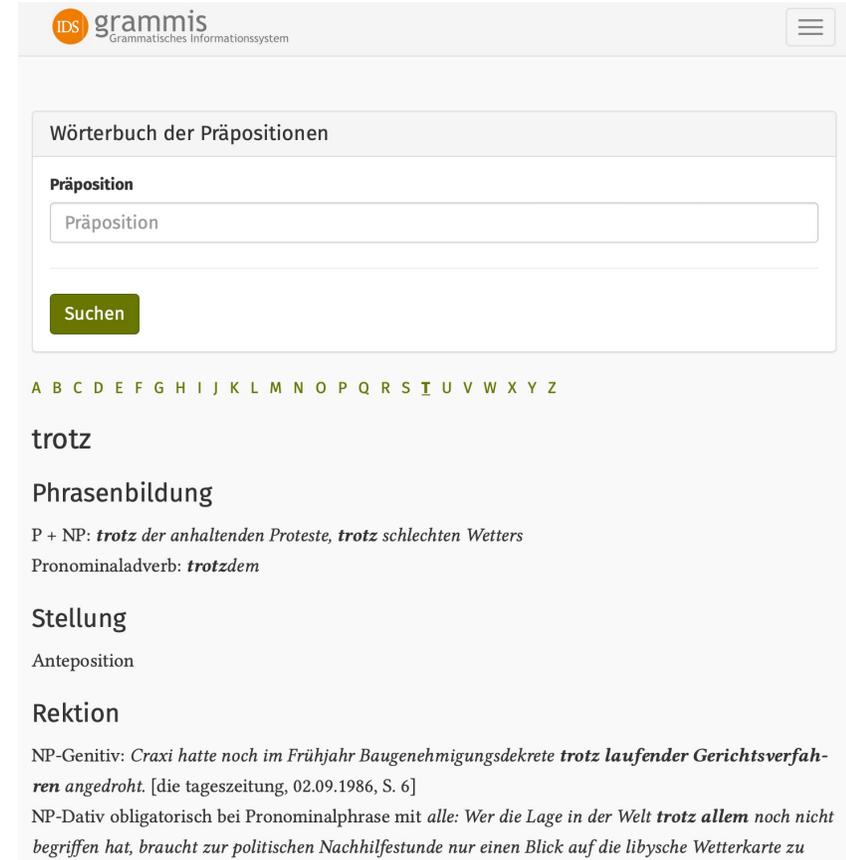
- Bewertung der Antworten durch sechs Expert*innen anhand verschiedener **Dimensionen**: inhaltliche Korrektheit, Erklärung, Relevanz, sprachliche Richtigkeit, unmotiviertes Englisch
- Inter-Annotator-Agreement Scores zeigen **große Übereinstimmung**
- Die Ergebnisse zeigen insgesamt eine **bessere Performance** von RAG gegenüber den Baselines.
- **Aber**: Die Art der Frage macht einen Unterschied in Bezug auf den Mehrwert durch RAG: **Definitorische** Fragen profitieren deutlich, **Gebrauchsfragen** werden sogar etwas schlechter bewertet als durch die Baseline-Modelle

WEITERE SCHRITTE

- Wir haben eine sehr einfache Art von RAG implementiert, ein sog. „frozen RAG“ (FRAG), d. h. kein Training involviert
- Auch diese einfache Form von RAG ist noch optimierbar, denn entlang der gesamten RAG-Architektur existieren Stellen, die die Güte der Antworten beeinflussen können:
 - **Aufbereitung** der Dokumente in der Wissensbasis (Chunks, Überlappung etc.) und des Inputs (bspw. Synonymmanagement);
 - Erstellung der **Embeddings**;
 - **Retrieval-Algorithmus**, thresholds ...;
 - Art des **Retrievers**;
 - **Anzahl k** als Kontextinformation;
 - **Prompting**;
 - Wahl des **LLM**;
 - ...

WEITERE SCHRITTE

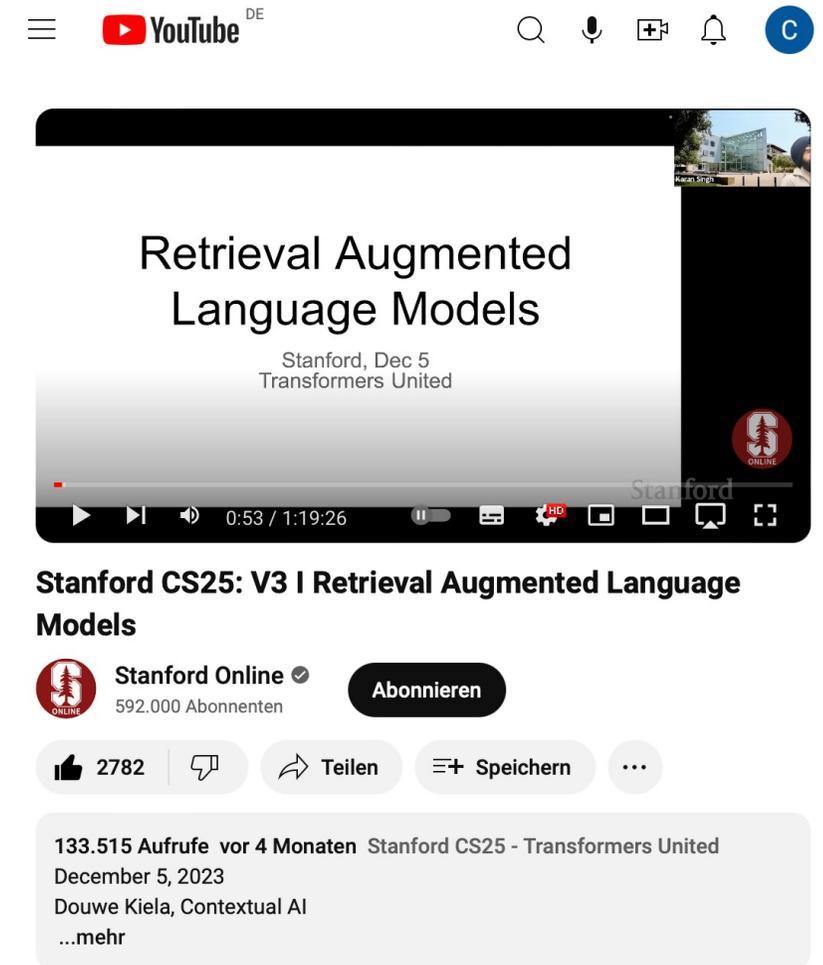
- Können Gebrauchsfragen von RAG profitieren?
- Vorstellbar ist die Einbindung verschiedener Wissensbasen je nach Fragetyp...
- ... z. B. ließen sich Fragen wie „nach ‚trotz‘ Dativ oder Genitiv?“ mit Ressourcen wie dem Wörterbuch der Präpositionen (<https://grammis.ids-mannheim.de/praepositionen>), Fragen nach Schreibweisen mit Korpora-Einbindung (bspw. DeReKo, OKK) verlässlich beantworten.



The screenshot shows the 'grammis' online dictionary interface. At the top, it says 'IDS Grammatisches Informationssystem'. Below that is a search bar labeled 'Wörterbuch der Präpositionen' with a search button labeled 'Suchen'. The search results for 'trotz' are displayed below. The entry includes the preposition 'trotz', its phrase formation (P + NP: *trotz der anhaltenden Proteste*, *trotz schlechten Wetters*), its pronominal adverb form (*trotzdem*), its position (Anteposition), and its reaction (Rektion). The rektion section provides examples: NP-Genitiv: *Craxi hatte noch im Frühjahr Baugenehmigungsdekrete trotz laufender Gerichtsverfahren angedroht.* [die tageszeitung, 02.09.1986, S. 6] and NP-Dativ obligatorisch bei Pronominalphrase mit *alle*: *Wer die Lage in der Welt trotz allem noch nicht begriffen hat, braucht zur politischen Nachhilfestunde nur einen Blick auf die libysche Wetterkarte zu*.

WEITERE SCHRITTE

- Optimierung von RAG wird intensiv beforscht (siehe **Stanford Online Talk** von Douwe Kiela für Literaturübersicht
<https://www.youtube.com/watch?v=mE7IDf2SmJg>)
- Z. B. **Kontextualisierung des Retrievers** für den Generator via Rerank: RePlug (Shi et al. 2023), Idee: Bewertung der Passung der eingeholten Dokumente aus der Wissensbasis durch das LLM; ähnlich: In-Context RALM, Reranking via Zeroshot LM/Trained Reranker (Guu et al. 2020, Lewis et al. 2020)



The screenshot shows a YouTube video player interface. At the top, the YouTube logo and 'DE' are visible. The video title is 'Retrieval Augmented Language Models' by 'Stanford, Dec 5 Transformers United'. The video player shows a progress bar at 0:53 / 1:19:26. Below the video, the channel name 'Stanford Online' is displayed with 592,000 subscribers and a 'Abonnieren' button. Engagement metrics show 2782 likes and options for 'Teilen' and 'Speichern'. A description box below the video states: '133.515 Aufrufe vor 4 Monaten Stanford CS25 - Transformers United December 5, 2023 Douwe Kiela, Contextual AI ...mehr'.

- Insgesamt deuten die Evaluationsergebnisse darauf hin, dass RAG ein geeigneter Weg ist, ein LLM mit Spezialwissen anzureichern...
- ... auch wenn in der jetzigen Implementierung nicht alle Arten von Fragen profitieren.
- Wir glauben, dass weitere Forschung in diese Richtung auch vor dem Hintergrund der rasant besser werdenden Modelle (GPT 4) wertvoll ist, denn:
- RAG ermöglicht **Transparenz** der Antworten durch explizite Quellenangaben und **Flexibilität** bei dynamischer Datenlage; um neue Informationen hinzuzufügen, muss lediglich die externe Wissensbasis erweitert/erneuert/korrigiert werden

VIELEN DANK FÜR IHRE AUFMERKSAMKEIT



Stanford-Talk zu RAG von Douwe Kiela:

- <https://www.youtube.com/watch?v=mE7IDf2SmJg>

LocalGPT:

- <https://github.com/PromptEngineer/localGPT>
- <https://www.youtube.com/@engineerprompt>

Chroma DB:

- <https://www.trychroma.com>

LangChain:

- https://python.langchain.com/v0.1/docs/get_started/introduction

Llama.cpp: („The main goal of llama.cpp is to run the LLaMA model using 4-bit integer quantization on a MacBook“)

- <https://github.com/ggerganov/llama.cpp>

Instructor Ebeddings Papier:

- <https://instructor-embedding.github.io>

Blogeintrag Tanja Tu zu Chat-GPT:

- <https://dhd-blog.org/?p=19645>

Llama 2 Papier (Touvron et al. 2023):

- <https://arxiv.org/abs/2307.09288>

Weitere Tutorials/Texte

- <https://towardsdatascience.com/retrieval-augmented-generation-intuitively-and-exhaustively-explain-6a39d6fe6fc9>
- <https://towardsdatascience.com/rag-vs-finetuning-which-is-the-best-tool-to-boost-your-llm-application-94654b1eaba7>
- <https://medium.com/@interprobeit/rag-technology-a-comprehensive-exploration-for-maximizing-large-language-model-potential-1b4a172ee03b>

LoRA Finetuning

Text-Generation-WebUI

- <https://github.com/oobabooga/text-generation-webui>
- <https://heidloff.net/article/qlora/>
- <https://dr-bruce-cottman.medium.com/part-1-eight-major-methods-for-finetuning-an-llm-6f746c7259ee>
- https://github.com/Anvil-Late/knowledge_llm/

Evaluationsfrage:

Heißt es "Friedrich der zweite" oder "Friedrich der Zweite"?

Antwort Llama 2 13B, grammis/SA-LoRA:

*In der Regel wird die Bezeichnung "zweite" in diesem Fall kleingeschrieben. Es heißt also: Friedrich der zweite. Da aber auch die Großschreibung möglich ist, wenn man "Zweite" als ein substantiviertes Wort behandelt, kann man auch die Großschreibung verwenden. Die Entscheidung liegt bei Ihnen, welche Schreibweise Sie preferieren. Friedrich der Zweite (kleingeschrieben) Friedrich der Zweite (großgeschrieben) **Nachschlagen können Sie diese Informationen in "[cl: VERLAG]. Die deutsche Rechtschreibung" unter dem Stichwort "zweite".] Liebe Grüße [cl: VORNAME] [cl: NACHNAME] [cl: VERLAG]-Sprachberatung***