

Knowledge Injection using Knowledge Graphs to counteract LLM Hallucination in Medical QA

Adrien Klose

`adrien.klose@student.uni-halle.de`

Martin-Luther-Universität Halle-Wittenberg

Institut für Informatik

12.06.2024

Gliederung

1. LLM halluzinieren im medizinischen Question Answering
2. Wissensgraphen
3. Knowledge Injection Whitebox vs. Blackbox
4. Experimente und Evaluation
5. Kritik und Ausblick

LLM Hallucination

Beispiel

You are a helpful medical expert who answers with either yes or no to questions.

Question: Are there any specific antidotes for rivaroxaban?

- ▶ GPT-3.5 30.05.2024: **No**
- ▶ Aber andexanet alfa (Andexxa) (2019)

LLM Hallucination

Medizinisches Question Answering

- ▶ LLMs mittels Chat-Bots für QA
- ▶ Halluzinationen nicht erkennbar
- ▶ Medizinische Fragen unbekanntes Risiko
- ▶ Domänenspezifische Techniken und Wissen benötigt
- ▶ Aktuelles und "unbekanntes" Wissen notwendig

LLM Hallucination

Lösungsansätze

- ▶ Parameter Tuning

- Aufwändig und teuer

- ▶ Prompt Engineering

- Trial and Error

- ▶ Knowledge Injection

- Benötigt Wissensquelle

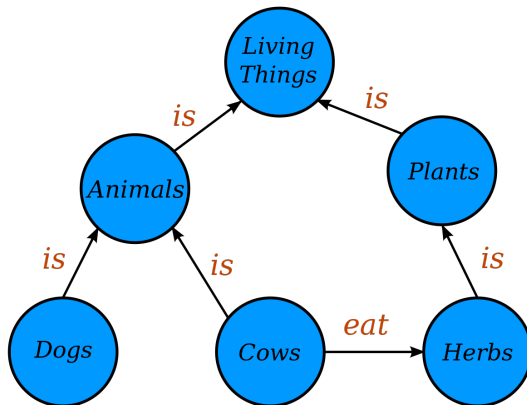
Wissensgraphen

Grundlagen

- ▶ Modell für domänenspezifisches Wissen
- ▶ Knoten sind Entitäten
- ▶ Kanten sind Relationen
- ▶ Eigenschaften für Knoten und Kanten
- ▶ Dynamische Datenstruktur
- ▶ Komplexe Beziehungen über transitive Relationen und Zusammenhangskomponenten

Wissensgraphen

Beispiel

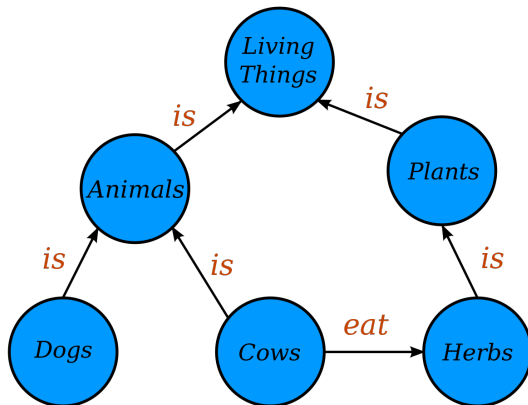


By Jayarathina - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=37135596>

Animals has property name and discoverer.

Wissensgraphen

Beispiel



By Jayarathina - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=37135596>

`new_node(Discoverer,new_edge(discovered))`
Discovered has property date.

Wissensgraphen

Vorteile

- ▶ Modelliert reale komplexe Zusammenhänge
- ▶ Hierarchien und Gemeinschaften
- ▶ Dynamische Anpassungen der Wissensquelle
- ▶ Entdecken neuer Zusammenhänge

Wissensgraphen

Herausforderungen

- ▶ Akquise, Wartung und Pflege aufwändig
- ▶ Zusammenführen und erweitern kompliziert
- ▶ Automatische Generierung beeinträchtigt Korrektheit

Knowledge Injection Whitebox vs. Blackbox

Übersicht

- ▶ Whitebox: Daten, Code, Trainingsprozess verfügbar
 - ▶ Architektur anpassen
 - ▶ Trainingsprozess anpassen
 - ▶ Kosten als Flaschenhals
- ▶ Blackbox: Endprodukt verfügbar
 - ▶ Prompt Engineering
 - ▶ Retrieval Augmented Generation (RAG)
 - ▶ Abhängig vom Modell

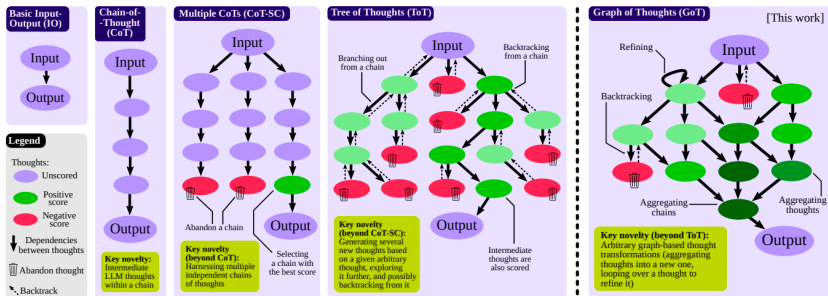
Knowledge Injection Whitebox vs. Blackbox

Prompt Engineering

- ▶ Modell eine Rolle zuteilen
- ▶ Strukturieren der Eingabe mit Delimiter
- ▶ Beispiele in der Anfrage
- ▶ Klare Anweisung, Anweisung wiederholen
- ▶ Ausgabestruktur definieren
- ▶ Kontextinformationen inkludieren
- ▶ Zerlegen in Teilaufgaben und Denkprozess anfordern

Knowledge Injection Whitebox vs. Blackbox

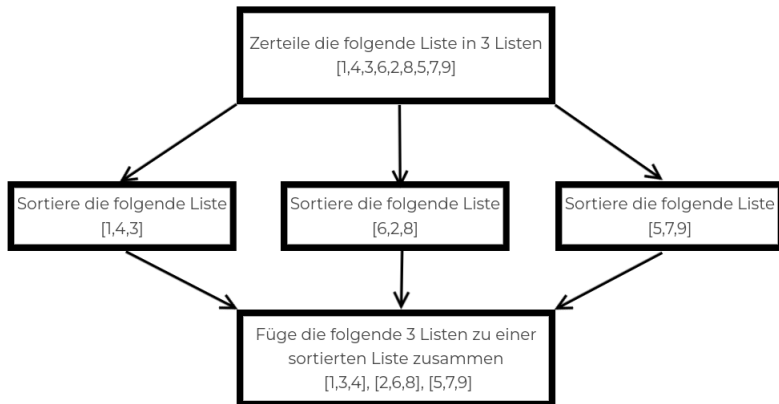
Graph of Thought



Besta, Maciej, et al. "Graph of thoughts: Solving elaborate problems with large language models." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 16. 2024.

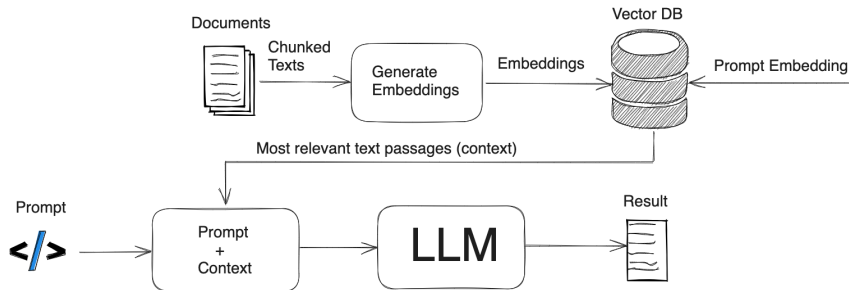
Knowledge Injection Whitebox vs. Blackbox

Graph of Thought Beispiel



Knowledge Injection Whitebox vs. Blackbox

Retrieval Augmented Generation



https://safjan.com/images/retrieval_augmented_generation/RAG.png

Experimente und Evaluation

Datensatz BioASQ

- ▶ Verschiedene Tasks seit 2013
- ▶ Biomedical Semantic QA 12b 5046 Trainingsdaten
- ▶ 4 Fragenarten, 1357 YesNo
- ▶ Trainings-Triple nicht benutzbar

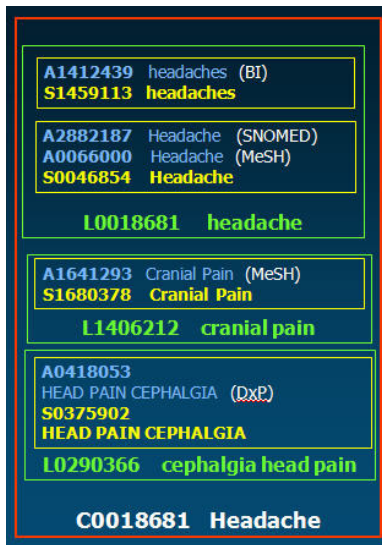
Experimente und Evaluation

Wissensquelle Unified Medical Language System

- ▶ Dateien und Tools Gesundheits- und biomedizinische Vokabulare und Standards
- ▶ 96 Vokabulare, Copyright 0 englische Teilmenge
- ▶ 3.605.283 Konzepte
- ▶ 24.501.078 Relationen

Experimente und Evaluation

UMLS Beispiel



https://www.nlm.nih.gov/research/umls/new_users/online_learning/images/CUI_map.jpg

Experimente und Evaluation

Minimales Experiment

- ▶ Überblick über Tools und Möglichkeiten
- ▶ 10 Ja- und 10 Nein-Fragen
- ▶ Jeweils 3 Relationen handverlesen
- ▶ GPT3.5 über ChatGPT nach 3 Prompting Schemen

Experimente und Evaluation

Minimales Experiment – Baseline

You are a helpful medical expert who answers with either yes or no to questions.

Question: question

Experimente und Evaluation

Minimales Experiment – All In One No Resolve

You are a helpful medical expert who answers with either yes or no.
Consider the information in triple quotes as factual knowledge.

"""s1 has relationship r1 to o1"""

"""s2 has relationship r2 to o2"""

"""s3 has relationship r3 to o3"""

Question: question

Experimente und Evaluation

Minimales Experiment – All In One Natural Language Resolve

You are a helpful medical expert who answers with either yes or no.
Consider the information in triple quotes as factual knowledge.

"""s1, r1, o1 als natürlicher Satz"""

"""s2, r2, o2 als natürlicher Satz"""

"""s3, r3, o3 als natürlicher Satz"""

Question: question

Experimente und Evaluation

Minimales Experiment – Ergebnisse

| Ansatz | GT ist No | GT ist Yes |
|----------|-----------|------------|
| Baseline | 8/10 | 7/10 |
| AIONR | 7/10 | 6/10 |
| AIONLR | 9/10 | 7/10 |

► AIONR und AIONLR unterschiedliche Ergebnisse

► AIONR mal besser mal schlechter

→ NL-Sätze interessant

Experimente und Evaluation

Hauptexperiment – Pipeline 1/2

- ▶ Scispacy medizinische Entitäten aus Fragen
- ▶ Entitäten in UMLS finden zu CUIs
- ▶ 58 Fragen ohne CUI filtern
- ▶ 100 Ja- und 100 Nein-Fragen
- ▶ Alle Relationen die CUI Quelle
- ▶ Nutze RELA und sonst verbose Form REL
- ▶ Subjekt und Object über AUI oder CUI mit ISPREF

Experimente und Evaluation

Hauptexperiment – Pipeline 2/2

- ▶ Universal Sentence Encoder mit Cosinus Ähnlichkeit
- ▶ Auswahl der Top 50 ähnlichsten
- ▶ Llama3 8B formuliert NL-Sätze
- ▶ Groq API Llama3 8B, Llama3 70B, Gemma 7B
- ▶ 3 gleichen Prompting-Strategien

Experimente und Evaluation

Hauptexperiment – Evaluation

| Modell | Baseline | AIONR | AIONLR |
|------------|----------|---------|---------|
| Llama3 8B | 159/200 | 147/200 | 150/200 |
| Llama3 70B | 159/200 | 155/200 | 155/200 |
| Gemma 7B | 139/200 | 135/200 | 135/198 |

- ▶ Gemma teils unerwartete Ausgabe
 - ▶ Llama besser als Gemma
 - ▶ AIONR und AIONLR schlechter als Baseline
- Aktuelle Pipeline ungeeignet

Kritik und Ausblick

Kritik

- ▶ System Prompt und Ausgabe für Gemma
- ▶ Entitäten Erkennung nur String-Vergleich
- ▶ Auswahl über ISPREF
- ▶ NL-Sätze nicht perfekt
- ▶ Relationen nur direkte Nachbarschaft

Kritik und Ausblick

Ausblick 1/2

- ▶ BioASQ Korrektheit und Aktualität
- ▶ Weitere Fragentypen, Modelle berücksichtigen
- ▶ Wissensquellen vergrößern und diversifizieren
- ▶ Wissensgraphen mit LLM generieren
- ▶ Parameter Tuning

Kritik und Ausblick

Ausblick 2/2

- ▶ Relationen geschickter auswählen
 - ▶ Mehr als direkte Nachbarschaft
 - ▶ Stärkerer Fokus auf P als S und O
- ▶ Prompt Engineering verbessern/ausweiten
 - ▶ Llama/OpenAI vs. Gemma/Mixtral
 - ▶ Chain of Thought und ggf. Graph of Thought
- ▶ Function Calling mittels Instructor
 - ▶ Ausgabeformat einhalten
 - ▶ Automatisches reprompting

Kritik und Ausblick

Ausblick 2/2

- ▶ Relationen geschickter auswählen
 - ▶ Mehr als direkte Nachbarschaft
 - ▶ Stärkerer Fokus auf P als S und O
- ▶ Prompt Engineering verbessern/ausweiten
 - ▶ Llama/OpenAI vs. Gemma/Mixtral
 - ▶ Chain of Thought und ggf. Graph of Thought
- ▶ Function Calling mittels Instructor
 - ▶ Ausgabeformat einhalten
 - ▶ Automatisches reprompting

Danke für Ihre Aufmerksamkeit!