

# **Wissensgraphen zur Verbesserung von LLMs in dem medizinischen Question Answering**

**Forschungsgruppenmodul Datenbanken und Informationssysteme**

Adrien Klose; akxeq

21. Oktober 2024

# 1 Einleitung

Large Language Models (LLMs) sind in dem Bereich der künstlichen Intelligenz eine junge Technologie, die ihren Ursprung in dem Aufmerksamkeitsmechanismus und der Transformer-Architektur hat [BCB14; Vas+17]. Seit der Veröffentlichung des ersten LLM BERT im Jahr 2018 von Google [Dev+18] wurden zahlreiche weitere und leistungstärkere LLMs von verschiedensten Organisationen wie GPT von OpenAI [Bro20], LLaMA von Meta [Tou+23] und BLOOM von hunderten verschiedenen Wissenschaftlern [Le+23] publiziert. Mit dem Wachstum von BERT mit hunderten Millionen Parametern zu heutigen LLMs mit hunderten Milliarden Parameter, die noch weiter wachsen, ändern sich fortlaufend die Dimensionen und Möglichkeiten von LLMs [Kum23].

Für Natural Language Processing Aufgaben wie automatisches Übersetzen, automatische Textzusammenfassung und Question Answering haben sich LLMs als eine erfolgreiche Technik erwiesen, die aus kaum einem State-of-the-Art Ansatz wegzudenken sind [Cha+24]. LLMs und ihre Fähigkeiten sind seit der Veröffentlichung von ChatGPT nicht nur für die Wissenschaft und Unternehmen zugänglich, sondern auch für die breite Bevölkerung [Ope24a]. ChatGPT mit seiner einfachen und kostenfreien Webschnittstelle erlaubt es die Möglichkeiten von LLMs für alltägliche Aufgaben zu verwenden. Insbesondere für die Beantwortung von Fragen als Alternative zur eigenen Recherche in verschiedenen Medien hat sich ChatGPT als ein Alltagshelfer herausgestellt [Mor24]. Die gestellten Fragen reichen von Übersetzungsanfragen über Textgenerationsanfragen bis hin zu Fragen über medizinische Probleme. Obwohl ChatGPT in der Regel korrekte und zufriedenstellende Antworten liefert, kommt es auch zu faktisch inkorrekten Antworten [Ji+23]. Ohne Wissen über die korrekte Antwort und weiteren Recherchen kann nicht unterschieden werden, ob die produzierten Antworten und Referenzen korrekt oder halluziniert sind [Ath+23; Bha+23].

Inkorrekte Antworten, die von Laien aber auch Experten kaum bis gar nicht erkannt werden können, sind insbesondere bei der Beantwortung von medizinischen Fragen ein großes Risiko. Eine inkorrekte Antwort auf eine medizinische Frage führt dazu, dass der Patient einem nicht einschätzbaren Risiko ausgesetzt ist. Damit LLMs mittels Chat-Bots wie ChatGPT besser im medizinischen Question Answering eingesetzt werden können, müssen LLMs um domänspezifische Techniken für komplexe medizinische Fachbegriffe in ihrem jeweiligen Kontext und faktisches Wissen ergänzt werden [Wan+23]. Die Möglichkeiten Wissen in LLMs zu inkludieren unterscheiden sich dabei stark, ob das LLM als eine Blackbox oder Whitebox zu betrachten ist [Pan+24]. LLMs als Whitebox erlauben es nicht nur Änderungen an der Architektur vorzunehmen, sondern auch den gesamten Trainingsprozess neu auszurichten [Liu+21b; Yas+21]. Diese Anpassungen bedeuten jedoch einen komplett neuen Trainingsprozess, der mit erheblichen Kosten verbunden ist. LLMs als Blackbox benötigen keinen extra Trainingsprozess aber erlauben es nur das existierende vortrainierte Modell zu erweitern oder die Eingabe geschickt zu manipulieren. Ohne die Architektur des LLMs um weitere Schichten und einem extra Trainingsprozess zu erweitern, kann aktuelles Wissen und Kontext dem LLM nur mittels Manipulation der Prompts und des gesamten Prompting Prozesses gegeben werden [HK23; Bes+24]. Die Techniken für LLMs als Blackbox können auch

im Whitebox-Szenario angewendet werden und sind insbesondere im Rahmen der Nutzung verschiedener Endnutzergruppen relevant, da ein endnutzerspezifisches bis hin zum personalisierten Training von LLMs kostentechnisch unrealistisch ist.

Für das medizinische Question Answering mittels Chat-Bots müssen aus den natürlichsprachlichen Anfragen die relevanten Entitäten extrahiert, mehrdeutige Terme aufgelöst, relevante kontextuelle Informationen identifiziert und diese dann geeignet dem Prompt hinzugefügt werden. Wissensgraphen als Modelle enkodieren domänenspezifisches Wissen über Entitäten und Relationen die zwischen Entitäten existieren. Im medizinischen Bereich existieren bereits eine Vielzahl an unterschiedlichen Wissensgraphen in unterschiedlichen Formaten [Jac+21; Bod04], jedoch müssen diese für den effektiven Einsatz vereinheitlicht und zusammengeführt werden.

In dieser Vorarbeit schaffen wir einen Überblick über das Unified Medical Language System (UMLS), welches als Wissensgraph viele verschiedene medizinische Vokabulare und Wissensgraphen zusammenführt [Bod04]. Wir benutzen dann einen Teil des UMLS zusammen mit verschiedenen LLMs, um den Einfluss auf die Genauigkeit der Beantwortung von Ja-Nein Fragen des BioASQ-Task [Nen+24] zu untersuchen. In unseren Experimenten stellen wir fest, dass wir mit unserem simplen Retrieval Augmented Generation Ansatz keine signifikante Verbesserung der Ergebnisse erzielen können. Ziel dieser Arbeit ist es eine generelle Richtung für die Verbesserung der Antworten von LLMs im medizinischen Question Answering mittels Wissensgraphen zu identifizieren.

Nachfolgend gehen wir genauer auf die einzelnen Schritte und damit verbundenen Probleme sowie bekannten Lösungsansätze im Question Answering ein. Anschließend geben wir einen Überblick über die Grundlagen von Wissensgraphen und Large Language Models. Danach charakterisieren wir die UMLS und BioASQ. Darauf gehen wir genauer auf unsere Experimente ein und diskutieren Verbesserungsmöglichkeiten. Zum Schluss geben wir einen Ausblick auf unsere zukünftige Arbeit.

## 2 Verwandte Literatur

### 2.1 Wissensgraphen

Wissensgraphen sind Modelle zum Organisieren von Wissen in einer Graphen ähnlichen Struktur [Hog+21]. Entitäten entsprechen den Knoten und Relationen zwischen den Entitäten entsprechen den Kanten im Graphen. Wissensgraphen als Modell erlauben gerichtete und ungerichtete Kanten und auch die Modellierung von Hypergraphen mit Hyperkanten. Sowohl die Kanten als auch Knoten können weitere Paare aus Eigenschaft und Wert besitzen. Insgesamt sind Wissensgraphen eine dynamische Datenstruktur für die nicht im vornherein ein festes Schema definiert werden muss. Die Flexibilität von Wissensgraphen erlaubt es diese dynamisch an neues Wissen und neue Erkenntnisse anzupassen, jedoch erschwert das Fehlen an Schematas die Interoperabilität verschiedener Wissensgraphen, die Wartung und Pflege von Wissensgraphen sowie das Stellen von Anfragen an den Graphen. Je nach Art und Komplexität des Graphen und den zu stellenden Anfragen muss aus Sprachen wie SPARQL [Har13], Cypher [Fra+18], Gremlin [Rod15] die geeignete ausgewählt werden. Insbesondere SPARQL als eine standardisierte Anfra-

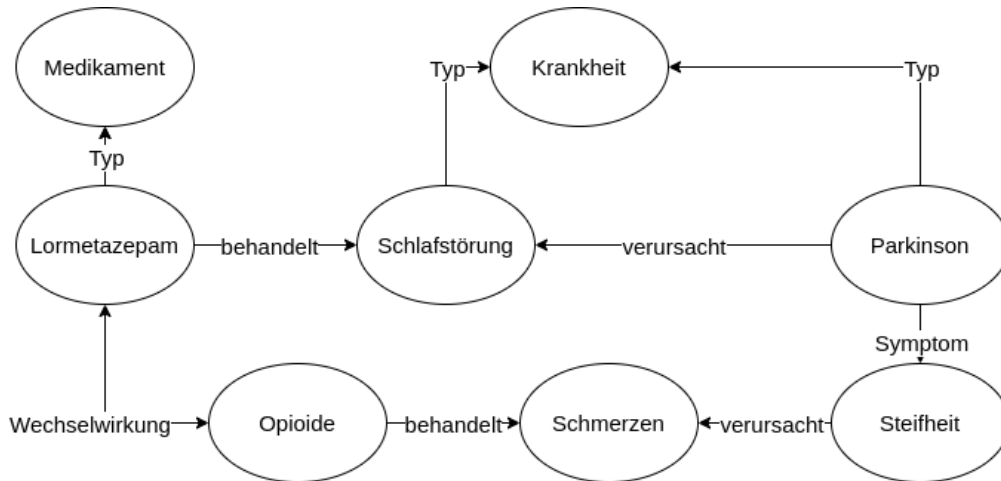


Abbildung 1: Beispiel für einen Wissensgraphen

gesprache für Tripel nach dem Resource Description Framework erlaubt es Anfragen an beliebige Datenstrukturen zu stellen, insofern diese sich in Tripel aus Subjekt, Prädikat und Objekt zerlegen lassen.

Die Nützlichkeit von Wissensgraphen hängt maßgeblich von ihrer Größe, Korrektheit und Vollständigkeit ab. Je nach Anwendungsszenario muss daher entschieden werden welche der Eigenschaften wie wichtig ist und der Wissensgraph derart erstellt und erweitert werden. HOGAN u. a. [HOG+20] unterteilt die Quellen für Wissensgraphen in manuell von Menschen, unstrukturierte und strukturierte Textquellen. Wissensgraphen können von Menschen unter anderem direkt durch Experten, Crowd-Sourcing, Feedback oder kollaborative Editing-Plattformen erstellt werden. Von Menschen erstellte Wissensgraphen weisen jedoch Herausforderungen auf bezüglich den Kosten der Erstellung, Bias, Lizenzierung und menschlichen Fehlern auf. Deswegen werden oftmals menschliche Experten nicht zur Erstellung von Wissensgraphen eingesetzt, sondern zur Verifizierung und Erweiterung von Wissensgraphen, die automatisch erstellt wurden. Für die automatische Erstellung von Wissensgraphen aus unstrukturierten Texten müssen die Texte mittels Vorverarbeitungsschritte wie Part-of-Speech tagging und Word Sense Disambiguation vorbereitet werden für die Named Entity Recognition und das Erkennen von Relationen zwischen den Entitäten. Nach der Identifikation relevanter Textabschnitte muss entschieden werden, ob und wie die extrahierten Relationen dem Wissensgraphen hinzugefügt werden, ob für die Entität bereits ein Synonymer Knoten existiert und ob sich aus den hinzugefügten Informationen neues Wissen ableiten lässt.

Anhand von Regeln und logischen Zusammenhängen kann deduktiv aus Wissensgraphen neues Wissen erzeugt werden. Aus dem abgebildeten Wissensgraphen 1 kann zum Beispiel abgeleitet werden, dass durch die Existenz der Verbindung von Parkinson zu Schmerzen über Steifheit auch Schmerzen ein indirektes Symptom für Parkinson sind. Für die Übertragbarkeit von Deduktionen und Zusammenhängen zwischen Wissensgraphen ist es hilfreich einer Ontologie als eine Konvention für bestimmte Bezeichnungen zur

Basis zu legen. Anhand von Ontologien zusammen mit dem Wissensgraphen und deduktiven Frameworks wie dem Rule Interchange Format, Datalog, Description Logics oder der Web Ontology Language 2 RL/RDF kann formalisiertes und transferierbares Wissen generiert werden [Hog+21]. Im Gegensatz zu deduziertem Wissen, das immer korrekt ist, ist induziertes Wissen nur zu einer bestimmten Wahrscheinlichkeit korrekt. Die Induktion für Wissen geschieht entweder überwacht oder unüberwacht. Mittels unüberwachter Methoden zur Bestimmung von Merkmalen wie dem PageRank, minimaler Schnitte und Zusammenhangskomponenten kann für jeden Wissensgraphen individuell induktives Wissen bestimmt werden, das je nach Kontext des Wissensgraphen unterschiedliche Aussagen erlaubt. In einem Wissensgraphen über Verkehrsnetzwerke können wir somit Aussagen über Flaschenhälse treffen und in einem Wissensgraphen über Medikamente könnten wir ähnlich wirkende Medikamente identifizieren. Überwachte Methoden erstelle eine latente Repräsentation des Wissensgraphen, anhand derer Problemstellungen probabilistisch beantwortet werden. Neben allgemeinerer Methoden des maschinellen Lernens wie Neuronale Modelle [BAH19; Det+18; Jaf+24], Sprachmodellen [Coc+17] und translatorischer Modelle die Regeln berücksichtigen [Guo+16] existieren auf Wissensgraphen ausgerichtete Graph Neural Networks [Sca+08].

## 2.2 Question Answering

Das Question Answering ist ein Teilgebiet des Natural Language Processing, in dem auf eine natürlichsprachliche Frage eine Antwort geliefert werden soll. Die Art der erwarteten Antwort variiert je nach Frage von Ja-Nein, über bestimmte Entitäten und Fakten bis hin zu komplexen Antworten [KM11]. Die Antworten können dabei aus einer vordefinierten Menge an Antwortmöglichkeiten stammen oder frei generiert werden. Gegeben eine Frage muss ein wissensbasiertes Question Answering System identifizieren welche Art von Antwort erwartet wird, welche Kandidaten für eine Antwort existieren und eine Antwort liefern [AH12]. Systeme basierend auf LLMs konsolidieren diese drei Schritte in das erstellen geeigneter Prompts aus Fragen und erzielen dabei vergleichbare Ergebnisse zu wissensbasierten State-of-the-Art Systemen [Tan+23]. Für die Erstellung geeigneter Prompts müssen analog zu den wissensbasierten Question Answering Systemen Entitäten erkannt, Mehrdeutigkeiten aufgelöst und kontextuelle Informationen identifiziert werden. Im Closed-Domain Question Answering wie in unserem Fall für die Medizin können dabei spezialisiertere Ansätze als beim Open-Domain Question Answering verfolgt werden.

Entitäten im medizinischen Kontext, die das gleiche Konzept beschreiben, haben eine große Varianz in der verwendeten Terminologie, die vom konkreten medizinischen Teilbereich abhängt [AZ11]. Zusammen mit ständig neu entstehenden medizinischen Fachbegriffen wie neuen Krankheiten und Medikamenten reichen Ansätze, die sich auf Wörterbücher beschränken, nicht aus für die Erkennung von Entitäten in der Medizin. Für die Erkennung medizinischer Entitäten müssen die Entitäten von anderen Textabschnitten abgegrenzt werden und anschließend die Art der Entität festgestellt werden. Ansätze sind danach zu unterscheiden, ob sie die Schritte simultan oder sequentiell ausführen und ob dies regelbasiert oder mittels maschinellern Lernen geschehen [AZ11;

Liu+21a]. Die gleiche Unterscheidung ist für Ansätze zur Auflösung von Mehrdeutigkeiten zu treffen. Neben wörterbuchbasierten Verfahren wie MetaMap [AL10] und cTAKES [Sav+10] werden auch Ansätze des maschinellen Lernens wie NCL [Vre+21] und kombinierte Ansätze [Kor+15] verwendet.

Aktuell lassen sich Ansätze für das medizinische Question Answering grundsätzlich in die Kategorien Information Retrieval, Question Entailment, Knowledge Base, Machine Reading Comprehension und Kombinationen daraus unterscheiden [Jin+22]. Information Retrieval basierte Ansätze [Liu13; Zhu+19] suchen nach relevanten Dokumenten und Passagen in Dokumenten, die direkt als Antwort auf die Frage ausgegeben werden. Für die Erstellung sinnvoller Antworten sind diese Systeme auf qualitative Datensätze an Dokumenten wie PubMed [Whi20] angewiesen. Question Entailment Ansätze [BD19] suchen zu gestellten Fragen ähnliche Fragen, deren Antworten bekannt sind und dessen Antwort auf die gestellte Frage übertragen werden können. Die beschränkte Menge an existierenden qualitativen Datensätzen an Frage-Frage und Frage-Antwort Paaren und die Kosten für das erstellen solcher Datensätze ist eine der Hauptherausforderungen von Question Entailment Ansätzen [Jin+22]. Knowledge Base Ansätze [Mar17] beantworten Fragen mittels Wissensbasen aus Entitäten und Relationen zwischen den Entitäten. Natürlichsprachliche Anfragen werden zunächst in SPARQL Anfragen umgewandelt, diese an die Wissensbasis gestellt und aus dem Ergebnis wird dann eine Antwort erstellt [Jin+22]. Die Güte der Antworten hängt dabei maßstäblich von der Qualität der Wissensbasen und der korrekten Erstellung der SPARQL Anfragen ab. Die UMLS ist dabei eine der wichtigsten Quellen für medizinische Daten die mehr als 200 verschiedene Wissensbasen zusammenfasst [Bod04]. Machine Reading Comprehension Ansätze [He+20; Jin+19; Per+24] werten die gegebene Frage zusammen mit einem gegebenen Kontext aus, um eine Antwort zu erstellen. Der Kontext für das Question Answering kann vorgegeben sein oder selber mittels Information Retrieval und Knowledge Base Ansätzen erstellt werden. Damit die Qualität der Antworten hoch ist müssen sowohl die Kontexte als auch Modelle, die Kontext und Frage zusammenführen, geeignet erstellt werden.

## 2.3 Large Language Models

Large Language Models sind Deep Learning Modelle, die auf dem Aufmerksamkeitsmechanismus und der Transformer-Architektur basieren, um auf natürlichsprachliche Anfragen mit natürlichsprachlichen Antworten zu reagieren. Dafür durchlaufen LLMs einen 3-stufigen Trainingsprozess aus Self Supervised Learning indem ein allgemeines Sprachverständnis trainiert wird, Supervised Learning indem ein Verständnis von Instruktionen und Interaktion trainiert wird und Reinforcement Learning indem trainiert wird, welche Antwortmuster zu bevorzugen sind [Liu+24]. Während dem Reinforcement Learning wird insbesondere darauf Wert gelegt, dass die Antworten von LLMs hilfreich, aufrichtig und harmlos sind. Im Angesicht der schieren Menge an benötigten Trainingsdaten ist es praktisch unmöglich bereits im vornherein Bias und falsche Informationen vollkommen zu eliminieren. Damit Halluzinationen, Bias und veraltete Antworten bewältigt und Antworten erklärbarer werden, müssen neben einem qualitativen Training noch weitere

Techniken angewendet werden.

Techniken zur Verbesserung der Antworten von LLMs lassen sich unterscheiden in Anpassungen am Trainingsprozess in Form von aufwändigerer Vorverarbeitung der Trainingsdaten und Finetuning auf spezifische Aufgaben und Domänen, und Anpassung an den Prompting-Prozess. Zhang u. a. [Zha+19] und Rosset u. a. [Ros+20] zeigten mit ihren Ansätzen wie Fakten aus Wissensgraphen aktiver in den Trainingsprozess mit einbezogen werden können, um wissensbasierte Aufgaben besser zu beantworten. Prompt Engineering hingegen erlaubt es ohne direkte Veränderungen am Modell oder zusätzliches Training die Antworten auf bestimmte Domänen und Aufgaben zu spezialisieren [Sah+24]. Few-Shot Prompting fügt den Anfragen Beispiele von erwarteten Ein- und Ausgabe-Paaren hinzu, um ein allgemeines Verständnis von der Aufgabe in der Anfrage zu inkludieren [Bro20]. Chain-of-Thought, Tree-of-Thought und Graph-of-Thought sind Techniken, die Modelle dazu bringen ihre Antworten in Form eines Argumentationsprozesses zu liefern [Bes+24]. Der Argumentationsprozess zerlegt Aufgaben in einfachere Teilschritte und fördert die Erklärbarkeit der Antworten. Eine weitere Möglichkeit den Argumentationsprozess zu nutzen ist Self-Consistency, bei dem zunächst verschiedene Argumentationsprozesse exploriert werden und anschließend der konsistenteste Argumentationsprozess verwendet wird [Wan+22]. Retrieval Augmented Generation ist eine Technik welche den Prompting-Prozess um traditionelles Information Retrieval erweitert [Lew+20]. Anhand der Anfrage werden in externen Wissensquellen wie Wissensgraphen oder Korpora relevante Informationen gesucht, geranked und geeignet der Anfrage hinzugefügt. Neben den allgemein anwendbaren Prompt-Engineering Techniken gibt es von den Herausgebern der Modelle wie ChatGPT und Llama auch auf ihr Modelle zugeschnittene Hinweise wie Prompts formuliert werden sollten [Ope24b; Met24].

In umgekehrter Reihenfolge wie LLMs Wissensgraphen für das Question Answering benutzen, werden LLMs verwendet, um Wissensgraphen zu erstellen, erweitern und Question Answering zu unterstützen [Pan+24]. Mittels LLMs können die Entitäten und Relationen mit ihren textuellen Beschreibungen aus Wissensgraphen in einen angereicherten Wortraum eingebettet werden und mit den Repräsentationen dann verschiedene Aufgaben erfüllt werden [Pan+24]. Neben der Möglichkeit zur Erstellung von Einbettungen werden LLMs auch verwendet, um gegeben zwei Teile eines Tripels direkt mögliche Kandidaten für die Vervollständigung des Tripels zu finden [Che+22]. Für die Erstellung von Wissensgraphen von Anfang an mittels LLMs können entweder die einzelnen Natural Language Processing Aufgaben von LLMs übernommen werden oder diese direkt aus dem latenten Wissen des LLMs erstellt werden. Bosselut u. a. [Bos+19] und Hao u. a. [Hao+22] schlagen verschiedene Techniken vor mittels derer aus einem kleinen anfänglichen Wissensgraphen automatisch größere Wissensgraphen erstellt werden. Die erstellten Wissensgraphen können dann zusammen mit LLMs zum Question Answering verwendet werden. LLMs werden verwendet zur Extraktion der Entitäten und Relationen aus den natürlichsprachlichen Anfragen [Hu+23], dem generieren von SPARQL Anfragen [Yan+23; Per+24] und der Filterung der Antwort Kandidaten [Yan+21].

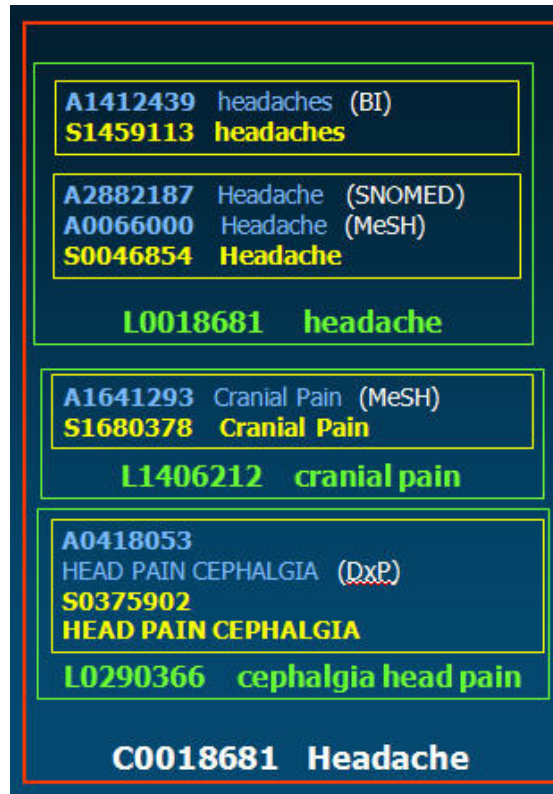


Abbildung 2: Beispiel für die Bezeichner-Struktur im Unified Medical Language System [Med16].

## 3 Experiment

### 3.1 Unified Medical Language System

Als Wissensgrundlage für unsere Experimente verwenden wir das Unified Medical Language System [Bod04]. Die UMLS ist eine Sammlung an Dateien und Software, welche verschiedene medizinische Vokabulare und Standards in einem Metathesaurus und einem semantischen Netzwerk zusammenführt. Zusätzlich zu englischen Vokabularen umfasst der Metathesaurus auch Vokabulare aus 17 anderen Sprachen. Mittels eindeutiger Bezeichner für Konzepte (CUI), Strings (SUI), lexikalische Begriffe (LUI), Relationen (RUI) und Atome (AUI) werden die Terme der Vokabulare hierarchisch strukturiert und voneinander unterscheidbar gemacht. Ein Konzept identifiziert eine Bedeutung, dass über mehrere verschiedene synonyme Terme beschrieben werden kann. Jedes Konzept hat mindestens 1 und maximal 5 zugewiesene Typen aus dem semantischen Netzwerk, wobei die möglichst spezifischsten semantischen Typen ausgewählt werden. Mittels Strings werden alle verschiedenen vorkommenden Zeichenkombinationen der Vokabulare zusammengefasst. Für jedes Konzept gibt es einen präferierten String. Lexikalische Begriffe fassen alle Strings zusammen, die mit dem Lexical Variant Generation Werkzeug des



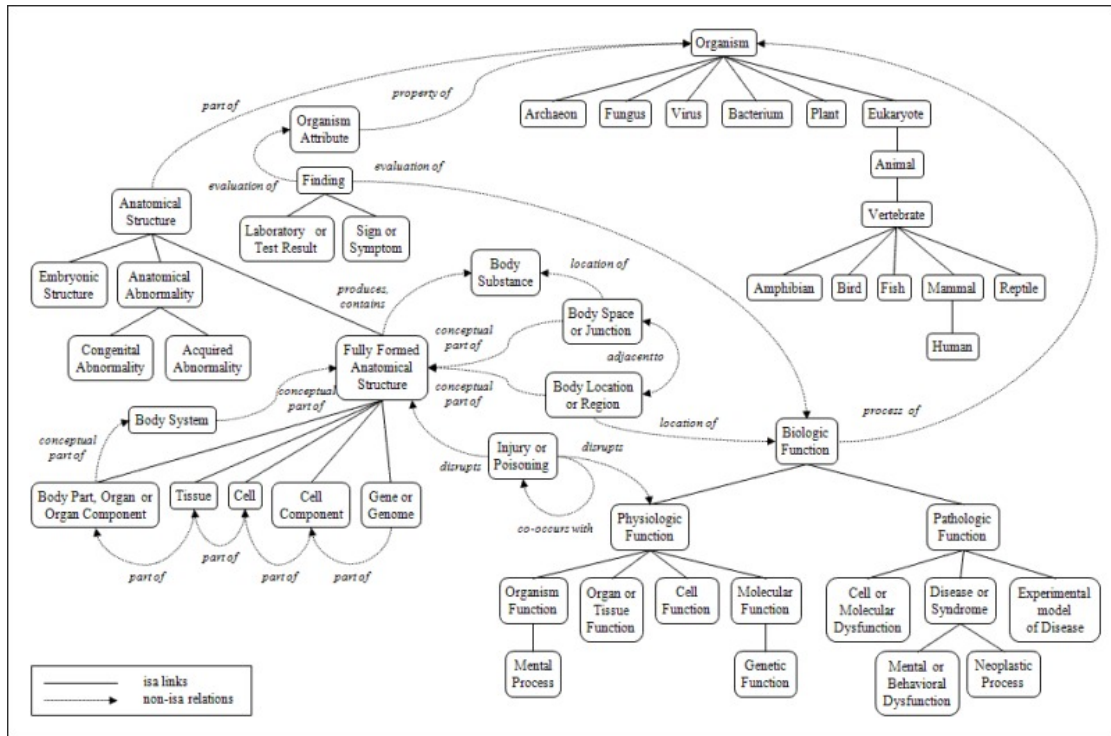


Abbildung 3: Ein Ausschnitt aus dem semantischen Netzwerk des UMLS [Med21].

UMLS auf die gleiche Grundform abgebildet werden. Atome sind die kleinste Einheit in der UMLS, wobei jedes verschiedene Vorkommen eines Begriffes in einem Vokabular seine eigene einzigartige Atom-Kennzeichnung bekommt. Ein Beispiel für die Struktur der Bezeichner anhand des Konzeptes Kopfschmerzen ist in Abbildung 2 zu sehen. Relationen zwischen Konzepten und Atomen werden mittels einer von 14 allgemeineren Relationen und optional einer von 1024 genaueren Relationen charakterisiert. Das semantische Netzwerk ergänzt die im Metathesaurus zusammengeführte Konzepte und Relationen um rund 130 verallgemeinerte semantische Typen mit rund 50 verschiedenen Relationen. Ein kleiner Ausschnitt aus dem semantischen Netzwerk ist in Abbildung 3 dargestellt. Die semantischen Typen werden unterschieden zwischen Entitäten wie Gene und Events wie mentale Prozesse. Für die Relationen im semantischen Netzwerk ist wichtig, dass diese sich nicht immer direkt auf das Level der Konzepte übertragen, so dass zum Beispiel Aspirin als ein Medikament nicht Krebs als eine Krankheit oder Syndrom auslösen kann. Die Vokabulare im UMLS unterliegen unterschiedlich strengen Levels an Lizenzvereinbarungen, sodass wir für unsere Experimente die vorgefertigte 2024AA Teilmenge an Vokabularen mit Lizenzlevel 0 verwenden. In dieser Teilmenge des UMLS sind 89 Vokabulare mit insgesamt 2.292.013 Konzepten und 20.091.016 Relationen enthalten.

## 3.2 BioASQ

Wir verwenden den BioASQ 12b Datensatz als Grundlage für das Question Answering [Nen+24]. BioASQ organisiert seit 2013 verschiedene Challenges in unter anderem Summarisation, Information Retrieval, Text-Klassifikation und Question Answering im biomedizinischen Bereich. Der 12b Datensatz enthält kombiniert 5046 Ja-Nein, Factoid, Listen und komplexe Fragen mit ihren korrekten Antworten für das Training. Die Trainingsdaten wurden seit 2013 angesammelt und von Experten zum jeweiligen Zeitpunkt beantwortet. Für 322 dieser Fragen existieren zusätzlich noch Tripel aus Subjekt, Prädikat und Objekt von ausgewählten Wissensquellen. Leider ist ein Teil der Wissensquellen nicht mehr zugreifbar, so dass auch die Tripel nicht mehr auswertbar sind. Deswegen berücksichtigen wir die Tripel nicht in unseren Experimenten. Für unsere Experimente beschränken wir uns auf 1357 Ja-Nein Fragen.

## 3.3 Versuchsaufbau

Wir untersuchen für unseren Ansatz zur Verbesserung von LLMs im medizinischen Question Answering mittels LLMs eine Anwendung des Retrieval Augmented Generation. Im ersten Schritt wählen wir aus den 1357 Ja-Nein Fragen des BioASQ 12b Datensatzes jeweils 100 Ja und 100 Nein Fragen aus. Wir exkludieren von der Auswahl 58 Fragen, für die wir mittels scyspaCy keine Entitäten erkennen konnten, die auch im UMLS vorkommen. Für jede gefundene Entität extrahieren wir aus der UMLS alle Relationen mit der Entität als Subjekt. Insofern vorhanden verwenden wir für Subjekt und Objekt den für das Atom angegebenen String und sonst verwenden wir den präferierten String des Konzeptes. Bei den Relationen verwenden wir die spezifischen Relationen und wenn diese nicht vorhanden sind, benutzen wir verbose Formen der allgemeinen Relationen. Anschließend erstellen wir mittels `dunzhang/stella_en_400M_v5` Einbettungen aus den Frage und den Tripeln aus Subjekt, Relation und Objekt. Mittels der Cosinus-Ähnlichkeit zwischen den Einbettungen erstellen wir ein Relevanz Ranking auf den Tripeln in Bezug zu der jeweiligen Frage. Anhand der Tripel und der Frage erstellen wir auf die folgenden 3 verschiedene Art und Weisen Anfragen an Llama3-8B, Llama3-70B und Gemma-7B und bestimmen die Genauigkeit bei einer Temperatur der Modelle von 0:

1. Als Baseline stellen wir den Modellen direkt die Frage und schränken die Antworten auf Ja oder Nein ein.

System: You are a helpful medical expert who must answer with either yes or no.

User: Are there any specific antidotes for rivaroxaban?

2. In diesem Ansatz fügen wir der Anfrage die bis zu 10 relevantesten Tripel als extra Informationen hinzu und stellen die Frage gefolgt von den Tripeln in einer Anfrage.

System: You are a helpful medical expert who must answer with either yes or no.

User: Are there any specific antidotes for rivaroxaban? Consider the following information:

```
s1 has relationship r1 to o1.
...
s10 has relationship r10 to o10.
```

- LLMs wurden konzipiert, um besser mit natürlicher Sprache zu interagieren. Deswegen lassen wir zunächst die 10 relevantesten Tripel mittels Llama3-8B in einen natürlichsprachlichen Text reformulieren,

System: You are a helpful writing assistant that can reformulate triples of subject, relationship and object into naturally sounding sentences. You only answer with the sentences.

User: Reformulate the following triples into one flowing text:

```
s1 p1 o1
...
s10 p10 o10
```

und fügen dann den reformulierten Text zu der Anfrage hinzu.

System: You are a helpful medical expert who must answer with either yes or no.

User: Are there any specific antidotes for rivaroxaban? Consider the following information:  
Reformulated sentences

Unsere Annahme ist, dass die reformulierten Tripel einen positiveren Einfluss auf die Genauigkeit als das bloße Beifügen der Tripel haben, da die natürlichsprachlich reformulierten Tripel näher an den erwarteten Eingabeformat der LLMs liegen sollte. Insbesondere bei den Fragen, die überwiegend Tripel mit allgemeinen Relationen hinzugefügt bekommen, erwarten wir stärkere Verbesserungen, da die verbosen Formen der allgemeinen Relationen zu unnatürlichen Satzformulierungen führen.

## 4 Evaluation

Gemäß unseren Ergebnissen in Tabelle 1 erzielen wir mit unserem Retrieval Augmented Generation Ansatz keine signifikant besseren Genauigkeiten gegenüber der Baseline. Für

Tabelle 1: Genauigkeit ausgewählter Modelle mit 3 verschiedenen Prompts auf 200 Fragen aufgeteilt nach der erwarteten Antwort. 100 Fragen erwarten Ja und 100 Fragen Nein als korrekte Antwort. Markierte Werte enthalten nicht Anfragen ohne Antworten.

Modell Ground Truth	Baseline		Tripel direkt		Tripel reformuliert	
	Ja	Nein	Ja	Nein	Ja	Nein
Llama3-8B	0.89	0.72	0.82	0.78	0.65	0.84
Llama3-70B	0.98	0.67	0.95	0.68	0.82	0.74
Gemma-7B	0.94	0.40	0.88*	0.43*	0.90*	0.41*

die Auswertung der Antworten haben wir bei Gemma ohne Reformulierung der Tripel 17 und bei Gemma mit Reformulierung der Tripel 40 Fragen ausgeschlossen. Bei jeder dieser Fragen konnte weder ein Ja noch ein Nein aus der Antwort extrahiert werden, da entweder die Antwort des Modelles daraus bestand, dass anhand der gegebenen Informationen keine Entscheidungen getroffen werden kann oder eine Liste aus Ja und Nein zurückgegeben wurde. Insgesamt bestehen die Antworten für die Llama-Modelle nur aus einem Ja oder Nein und die Antworten für Gemma enthielten neben Ja oder Nein mehrheitlich noch einen Nachsatz, der die gegebene Antwort mittels der gegebenen Informationen versucht zu begründen. Sowohl die Begründungen, die erzielten Ergebnisse als auch ein genauere Betrachtung der nach Relevanz sortierten Tripel zeigt, dass für die Verbesserung der Ergebnisse der gesamte Versuchsaufbau überarbeitet werden muss. In der aktuellen Iterationen erzeugen die hinzugefügten Informationen eher einen Topic Drift und lenken stärker von der gestellten Frage ab, als dass die Informationen dem Modell bei der Beantwortung der Frage helfen.

Im ersten Schritt der Bestimmung der Entitäten können wir zwar mit scyspaCy neben 1-Wort Entitäten auch zusammengesetzte Entitäten wie Hirschhaus Syndrome erkennen, jedoch erhalten wir ebenfalls viele False Positivs und False Negatives. Bei den False Positives stechen insbesondere als Entitäten nicht nützliche Verben wie **decrease**, **increase**, **improve**, **triggered** und zu allgemeine Substantive wie **outcomes**, **risk** hervor. Die Auswirkungen der False Negatives sind gravierender als die der False Positives, da Entitäten wie **subacute thyroiditis**, **amyotrophic lateral sclerosis** entweder garnicht oder nicht zusammenhängend als Entitäten erkannt wurden. Während die False Positives zum Aufblähen der Anzahl der gefundenen Tripel führt, können durch die False Negatives relevante Informationen und Tripel garnicht erst identifiziert werden. Neben der Minderung der False Negatives sollten wir die False Positive Verben ausfiltern und ausgiebiger benutzen, um zu bestimmen, welche Relationen zwischen den Entitäten für die Beantwortung der Frage nützlich sind.

Im zweiten Schritt haben wir die direkte Nachbarschaft jeder Entität als mögliche Kandidaten für Tripel betrachtet und diese nach der semantischen Ähnlichkeit zu der Anfrage als Indikator für die Relevanz sortiert. Die direkte Nachbarschaft zusammen mit der semantischen Ähnlichkeit führt zu dem Problem, dass unter den als sehr relevant eingeschätzten Tripeln Synonym-Relationen häufiger vorkommen. In unserem Anwendungsfall sind die Synonym-Relationen nicht informativ genug, um die Antworten der Modelle zu verbessern. Möglichkeiten, informativere Tripel zu erhalten, sind bereits bei der Entitäten Bestimmung die Kandidaten für relevante Relationen einzuschränken, nur Tripel die einen Schwellwert bezüglich der Relevanz übersteigen zu berücksichtigen oder größere Nachbarschaften in dem Wissensgraphen zu untersuchen. Für größere Nachbarschaften können wir entweder direkt nach Pfaden bestimmter Länge zwischen Entitäten suchen oder automatisch aus den Fragen SPARQL Anfragen erstellen lassen. Bei sämtlichen Anpassungen müssen wir jedoch aus Sicht der Praktikabilität beachten, dass die optimierte Bestimmung der hinzuzufügenden Tripel nicht wie aktuell im Minuten sondern im Millisekunden bis maximal Sekunden Bereich liegen sollte.

Im letzten Schritt erstellten wir aus den Fragen und Tripeln Anfragen für die verschiedenen Modelle. Den erheblichen Unterschied in dem Format der Antworten zwischen

Llama und Gemma führen wir darauf zurück, dass Gemma die Rolle System nicht unterstützt. Grundlegende Instruktionen, die wir dem Llama Modellen mittels den System Teil der Anfrage liefern, müssen wir Gemma zukünftig auf angepasste Art und Weise zuführen. Die Einhaltung bestimmter Antwortformate können wir mittels angepasster Prompts, Function Calling und automatischem Reprompting unterstützen. Eine Anpassung der Prompts ist auch dahingehend sinnvoll, dass wir deutlicher hervorheben welcher Teil der Anfrage am relevantesten ist durch zum Beispiel wiederholen der Frage am Ende der Anfrage und wir spezifischer hinweisen, dass die Antwort nicht ausschließlich auf den mitgegebenen Tripeln basieren soll. Ebenfalls müssen wir die Qualität der reformulierten Tripel dahingehen erhöhen, dass diese nicht zu einem einzelnen Satz aus mehreren Haupt- und Nebensätzen reformuliert werden.

Für die Verbesserung der Ergebnisse sind auch Anpassungen an dem BioASQ Datensatz und dem UMLS Wissensgraphen notwendig. Der BioASQ Datensatz, wie wir ihn verwendet haben, hat das Problem, dass die Ground Truth sich für Fragen vereinzelt geändert haben. Zum Beispiel die Frage **Are there any specific antidotes for rivaroxaban?** hat als nicht zeitgemäße Ground Truth **No**, obwohl mittlerweile Adexanet alfa als Gegenmittel existiert. Zukünftig sollten wir mindestens den Test-Datensatz auf Korrektheit überprüfen. Für die UMLS ist wichtig, dass wir diese um Vokabulare anderer Lizenzlevel ergänzen und neben der UMLS ebenfalls weitere Wissensgraphen heranziehen. Mittels mehrerer Wissensgraphen könnten wir dann je nach Frage nur auf das Themengebiet der Frage spezialisierte Wissensgraphen verwenden oder Mehrheitsentscheidungen über die Kandidaten der Wissensgraphen durchführen.

## 5 Fazit

In dieser Vorarbeit haben wir einen Überblick über die Themen Question Answering und Large Language Models im Zusammenhang mit Wissensgraphen gegeben. Anschließend haben wir einen Einstieg in das Unified Medical Language System mit seiner Bezeichner-Struktur sowie den verschiedenen Vokabularen und dem semantischen Netzwerk gegeben. Danach betrachteten wir kurz die Historie und Kerndaten von BioASQ bezogen auf den Question Answering Task. Nachfolgend führten wir verschiedene Experimente basierend auf Retrieval Augmented Generation durch, um die Genauigkeit bei der Beantwortung von Ja-Nein Fragen mittels LLMs zu verbessern. Wir haben gezeigt, dass unser simpler Ansatz ungeeignet ist, um die Genauigkeit zu erhöhen und sogar Tendenzen zur Verschlechterung der Genauigkeit zeigt. Daraufhin haben wir verschiedene Schwachstellen an unserem Versuchsaufbau diskutiert und mögliche erwartete Verbesserungsvorschläge für jeden einzelnen Teilschritt des Experimentes dargestellt.

In unseren zukünftigen Arbeit werden wir neben Ja-Nein Fragen auch die weiteren Fragetypen Entitäten, Liste an Entitäten und komplexe Fragen untersuchen. Wir werden die verschiedenen vorgeschlagenen Verbesserungen an unserem Retrieval Augmented Generation Ansatz untersuchen und weitere Ansätze zum injizieren von Wissen aus Wissensgraphen in LLMs als Blackbox und Whitebox explorieren. Als Wissensquelle werden wir neben der UMLS noch weitere medizinische Wissensgraphen untersuchen,

zusammenführen und verwenden und für BioASQ werden wir die Korrektheit und Aktualität der Testdaten sicherstellen. Wir werden einen besonderen Fokus darauf legen wie wir Wissen mittels Wissensgraphen in kleine und mittelgroße LLMs inkludieren können, damit diese kompetitive Ergebnisse zu große LLMs erzielen. Abschließend werden wir betrachten, ob sich das Question Answering Problem dahingehend modifizieren lässt, dass ein Entscheider zunächst feststellt, ob das LLM überhaupt zusätzliche Informationen für die Beantwortung der Frage benötigt. Mit der Verbesserung kleinerer LLMs und dem Entscheider erhoffen wir uns nicht nur kleineren Institutionen die Fähigkeiten von größeren LLMs zu kleineren Preisen zur Verfügung zu stellen, sondern auch die Ergebnisse von größeren LLMs im medizinischen Question Answering zu verbessern.

# Literatur

- [Bod04] Olivier Bodenreider. “The unified medical language system (UMLS): integrating biomedical terminology”. In: *Nucleic acids research* 32.suppl\_1 (2004), S. D267–D270.
- [Sca+08] Franco Scarselli u. a. “The graph neural network model”. In: *IEEE transactions on neural networks* 20.1 (2008), S. 61–80.
- [AL10] Alan R Aronson und François-Michel Lang. “An overview of MetaMap: historical perspective and recent advances”. In: *Journal of the American Medical Informatics Association* 17.3 (2010), S. 229–236.
- [Sav+10] Guergana K Savova u. a. “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications”. In: *Journal of the American Medical Informatics Association* 17.5 (2010), S. 507–513.
- [AZ11] Asma Ben Abacha und Pierre Zweigenbaum. “Medical entity recognition: a comparaison of semantic and statistical methods”. In: *Proceedings of BioNLP 2011 workshop*. 2011, S. 56–64.
- [KM11] Oleksandr Kolomiyets und Marie-Francine Moens. “A survey on question answering technology from an information retrieval perspective”. In: *Information Sciences* 181.24 (2011), S. 5412–5434.
- [AH12] Ali Mohamed Nabil Allam und Mohamed Hassan Haggag. “The question answering systems: A survey”. In: *International Journal of Research and Reviews in Information Sciences (IJRRIS)* 2.3 (2012).
- [Har13] S Harris. “SPARQL 1. 1 Query Language”. In: *W3C Recommendation* 21 (2013).
- [Liu13] Yifeng Liu. “The University of Alberta participation in the BioASQ challenge: The wishart system”. In: *Proc. 1st Workshop Bio-Med. Semantic Indexing Question Answering, Conf. Labs Eval. Forum*. 2013, S. 1–4.
- [BCB14] Dzmitry Bahdanau, Kyunghyun Cho und Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. In: *arXiv preprint arXiv:1409.0473* (2014).
- [Kor+15] Ioannis Korkontzelos u. a. “Boosting drug named entity recognition using an aggregate classifier”. In: *Artificial intelligence in medicine* 65.2 (2015), S. 145–153.

- [Rod15] Marko A Rodriguez. “The gremlin graph traversal machine and language (invited talk)”. In: *Proceedings of the 15th Symposium on Database Programming Languages*. 2015, S. 1–10.
- [Guo+16] Shu Guo u. a. “Jointly embedding knowledge graphs and logical rules”. In: *Proceedings of the 2016 conference on empirical methods in natural language processing*. 2016, S. 192–202.
- [Med16] National Library of Medicine (US). *Unique Identifiers in the Metathesaurus*. 2016. URL: [https://www.nlm.nih.gov/research/umls/new\\_users/online\\_learning/images/CUI\\_map.jpg](https://www.nlm.nih.gov/research/umls/new_users/online_learning/images/CUI_map.jpg) (besucht am 16.10.2024).
- [Coc+17] Michael Cochez u. a. “Global RDF vector space embeddings”. In: *The Semantic Web–ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part I 16*. Springer. 2017, S. 190–207.
- [Mar17] Anca Marginean. “Question answering over biomedical linked data with grammatical framework”. In: *Semantic Web 8.4 (2017)*, S. 565–580.
- [Vas+17] Ashish Vaswani u. a. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
- [Det+18] Tim Dettmers u. a. “Convolutional 2d knowledge graph embeddings”. In: *Proceedings of the AAAI conference on artificial intelligence*. Bd. 32. 1. 2018.
- [Dev+18] Jacob Devlin u. a. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).
- [Fra+18] Nadime Francis u. a. “Cypher: An evolving query language for property graphs”. In: *Proceedings of the 2018 international conference on management of data*. 2018, S. 1433–1445.
- [BAH19] Ivana Balažević, Carl Allen und Timothy M Hospedales. “Hypernetwork knowledge graph embeddings”. In: *Artificial Neural Networks and Machine Learning–ICANN 2019: Workshop and Special Sessions: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings 28*. Springer. 2019, S. 553–565.
- [BD19] Asma Ben Abacha und Dina Demner-Fushman. “A question-entailment approach to question answering”. In: *BMC bioinformatics* 20 (2019), S. 1–23.
- [Bos+19] Antoine Bosselut u. a. “COMET: Commonsense transformers for automatic knowledge graph construction”. In: *arXiv preprint arXiv:1906.05317* (2019).
- [Jin+19] Qiao Jin u. a. “Pubmedqa: A dataset for biomedical research question answering”. In: *arXiv preprint arXiv:1909.06146* (2019).
- [Zha+19] Zhengyan Zhang u. a. “ERNIE: Enhanced language representation with informative entities”. In: *arXiv preprint arXiv:1905.07129* (2019).



- [Zhu+19] Ming Zhu u. a. “A hierarchical attention retrieval model for healthcare question answering”. In: *The World Wide Web Conference*. 2019, S. 2472–2482.
- [Bro20] Tom B Brown. “Language models are few-shot learners”. In: *arXiv preprint arXiv:2005.14165* (2020).
- [He+20] Yun He u. a. “Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition”. In: *arXiv preprint arXiv:2010.03746* (2020).
- [HOG+20] AIDAN HOGAN u. a. “Knowledge Graphs”. In: *arXiv preprint arXiv:2003.02320* (2020).
- [Lew+20] Patrick Lewis u. a. “Retrieval-augmented generation for knowledge-intensive nlp tasks”. In: *Advances in Neural Information Processing Systems* 33 (2020), S. 9459–9474.
- [Ros+20] Corby Rosset u. a. “Knowledge-aware language model pretraining”. In: *arXiv preprint arXiv:2007.00655* (2020).
- [Whi20] Jacob White. “PubMed 2.0”. In: *Medical reference services quarterly* 39.4 (2020), S. 382–387.
- [Hog+21] Aidan Hogan u. a. “Knowledge graphs”. In: *ACM Computing Surveys (Csur)* 54.4 (2021), S. 1–37.
- [Jac+21] Rebecca Jackson u. a. “OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies”. In: *Database* 2021 (Okt. 2021), baab069. ISSN: 1758-0463. DOI: 10.1093/database/baab069. eprint: <https://academic.oup.com/database/article-pdf/doi/10.1093/database/baab069/40854912/baab069.pdf>. URL: <https://doi.org/10.1093/database/baab069>.
- [Liu+21a] Ning Liu u. a. “Med-BERT: A pretraining framework for medical records named entity recognition”. In: *IEEE Transactions on Industrial Informatics* 18.8 (2021), S. 5600–5608.
- [Liu+21b] Ye Liu u. a. “Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning”. In: *Proceedings of the AAAI conference on artificial intelligence*. Bd. 35. 7. 2021, S. 6418–6425.
- [Med21] National Library of Medicine (US). *UMLS® Reference Manual [Internet]*. 2021. URL: <https://www.ncbi.nlm.nih.gov/books/NBK9679/figure/ch05.F3/?report=objectonly> (besucht am 16.10.2024).
- [Vre+21] Alina Vretinaris u. a. “Medical entity disambiguation using graph neural networks”. In: *Proceedings of the 2021 international conference on management of data*. 2021, S. 2310–2318.
- [Yan+21] Yuanmeng Yan u. a. “Large-scale relation learning for question answering over knowledge bases with pre-trained language models”. In: *Proceedings of the 2021 conference on empirical methods in natural language processing*. 2021, S. 3653–3660.

- [Yas+21] Michihiro Yasunaga u. a. “QA-GNN: Reasoning with language models and knowledge graphs for question answering”. In: *arXiv preprint arXiv:2104.06378* (2021).
- [Che+22] Chen Chen u. a. “Knowledge is flat: A seq2seq generative framework for various knowledge graph completion”. In: *arXiv preprint arXiv:2209.07299* (2022).
- [Hao+22] Shibo Hao u. a. “BertNet: Harvesting knowledge graphs with arbitrary relations from pretrained language models”. In: *arXiv preprint arXiv:2206.14268* (2022).
- [Jin+22] Qiao Jin u. a. “Biomedical question answering: a survey of approaches and challenges”. In: *ACM Computing Surveys (CSUR)* 55.2 (2022), S. 1–36.
- [Wan+22] Xuezhi Wang u. a. “Self-consistency improves chain of thought reasoning in language models”. In: *arXiv preprint arXiv:2203.11171* (2022).
- [Ath+23] Sai Anirudh Athaluri u. a. “Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references”. In: *Cureus* 15.4 (2023).
- [Bha+23] Mehul Bhattacharyya u. a. “High rates of fabricated and inaccurate references in ChatGPT-generated medical content”. In: *Cureus* 15.5 (2023).
- [HK23] Thomas F Heston und Charya Khun. “Prompt engineering in medical education”. In: *International Medical Education* 2.3 (2023), S. 198–205.
- [Hu+23] Nan Hu u. a. “An empirical study of pre-trained language models in simple knowledge graph question answering”. In: *World Wide Web* 26.5 (2023), S. 2855–2886.
- [Ji+23] Ziwei Ji u. a. “Survey of hallucination in natural language generation”. In: *ACM Computing Surveys* 55.12 (2023), S. 1–38.
- [Kum23] Priyanka Kumari. *Comparing The Top Large Language Models For Multiple Use Cases*. 2023. URL: <https://www.labellerr.com/blog/comparing-language-models-through-parameters-vs-real-life-experiments/> (besucht am 21.05.2024).
- [Le +23] Teven Le Scao u. a. “Bloom: A 176b-parameter open-access multilingual language model”. In: (2023).
- [Tan+23] Yiming Tan u. a. “Can ChatGPT replace traditional KBQA models? An in-depth analysis of the question answering performance of the GPT LLM family”. In: *International Semantic Web Conference*. Springer. 2023, S. 348–367.
- [Tou+23] Hugo Touvron u. a. “Llama: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971* (2023).
- [Wan+23] Jindong Wang u. a. “On the robustness of chatgpt: An adversarial and out-of-distribution perspective”. In: *arXiv preprint arXiv:2302.12095* (2023).

- [Yan+23] Shuangtao Yang u. a. “Llm-based sparql generation with selected schema from large scale knowledge base”. In: *China Conference on Knowledge Graph and Semantic Computing*. Springer. 2023, S. 304–316.
- [Bes+24] Maciej Besta u. a. “Graph of thoughts: Solving elaborate problems with large language models”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Bd. 38. 16. 2024, S. 17682–17690.
- [Cha+24] Yupeng Chang u. a. “A survey on evaluation of large language models”. In: *ACM Transactions on Intelligent Systems and Technology* 15.3 (2024), S. 1–45.
- [Jaf+24] Parastoo Jafarzadeh u. a. “A Knowledge Graph Embedding Model for Answering Factoid Entity Questions”. In: *ACM Transactions on Information Systems* (2024).
- [Liu+24] Yiheng Liu u. a. “Understanding llms: A comprehensive overview from training to inference”. In: *arXiv preprint arXiv:2401.02038* (2024).
- [Met24] Meta. *Prompting*. 2024. URL: <https://www.llama.com/docs/how-to-guides/prompting/> (besucht am 10.10.2024).
- [Mor24] Oskar Mortensen. *How Many Users Does ChatGPT Have? Statistics & Facts (2024)*. 2024. URL: <https://seo.ai/blog/how-many-users-does-chatgpt-have> (besucht am 21.05.2024).
- [Nen+24] Anastasios Nentidis u. a. “BioASQ at CLEF2024: The Twelfth Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge”. In: *European Conference on Information Retrieval*. Springer. 2024, S. 490–497.
- [Ope24a] OpenAI. *Introducing GPT-4o and more tools to ChatGPT free users*. 2024. URL: <https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/> (besucht am 21.05.2024).
- [Ope24b] OpenAI. *Prompt engineering*. 2024. URL: <https://platform.openai.com/docs/guides/prompt-engineering/strategy-give-models-time-to-think> (besucht am 10.10.2024).
- [Pan+24] Shirui Pan u. a. “Unifying large language models and knowledge graphs: A roadmap”. In: *IEEE Transactions on Knowledge and Data Engineering* (2024).
- [Per+24] Aleksandr Perevalov u. a. “Language Models as SPARQL Query Filtering for Improving the Quality of Multilingual Question Answering over Knowledge Graphs”. In: *International Conference on Web Engineering*. Springer. 2024, S. 3–18.
- [Sah+24] Pranab Sahoo u. a. “A systematic survey of prompt engineering in large language models: Techniques and applications”. In: *arXiv preprint arXiv:2402.07927* (2024).