HAW HAMBURG

BACHELOR THESIS
Inken Dulige

Large Language Models im Gesundheitsbereich - Explorative Untersuchung freier Modelle für Vor-Anamnese

FAKULTÄT TECHNIK UND INFORMATIK
Department Informatik

Faculty of Engineering and Computer Science
Department Computer Science

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences

Inken Dulige

Large Language Models im Gesundheitsbereich - Explorative Untersuchung freier Modelle für Vor-Anamnese

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung
im Studiengang *Bachelor of Science Angewandte Informatik*
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr.-Ing. Christian Lins
Zweitgutachter: Prof. Dr. Philipp Jenke

Eingereicht am: 22.12.2023

Inken Dulige

Thema der Arbeit

Large Language Models im Gesundheitsbereich - Explorative Untersuchung freier Modelle für Vor-Anamnese

Stichworte

LLM, KI, Gesundheit, Vor-Anamnese

Kurzzusammenfassung

In dieser explorativen Studie soll untersucht werden, ob LLMs für die Erstellung einer Vor-Anamnese eingesetzt werden können. Dafür wurden Kriterien entwickelt, welche im Anschluss mit den passenden Benchmarks evaluiert werden sollen. Da die existierenden Benchmarks für die Evaluation nicht geeignet sind, kann keine fundierte Auswahl getroffen werden aber es besteht das Potenzial, dass LLMs zukünftig im Bereich der Vor-Anamnese eingesetzt werden können. ...

Inken Dulige

Title of Thesis

Large Language Models in the health sector - Explorative investigation of free models for pre-anamnesis

Keywords

LLM, AI, Healthcare, Pre-Anamnesis

Abstract

The aim of this exploratory study is to investigate whether LLMs can be used to take a preliminary medical history. Criteria were developed for this purpose, which will subsequently be evaluated using suitable benchmarks. As the existing benchmarks are not suitable for evaluation, it is not possible to make a well-founded selection, but there is potential for LLMs to be used in the field of pre-history taking in the future. ...

Inhaltsverzeichnis

Tabellenverzeichnis

1 Einleitung

Während der Corona-Pandemie haben das Zeitmanagement in Arztpraxen und die Möglichkeit der digitalen Unterstützung wie die telefonische und digitale Krankschreibung an medialer Aufmerksamkeit gewonnen. Gleichzeitig ist auch die Akzeptanz für „(...) die neuen intelligenten Wege der ambulanten medizinischen Versorgung (...)“ (Storm, 2021) [76] gestiegen. Bei den Beschäftigten im Gesundheitssystem wächst darüber hinaus das Interesse aus „(...) der Erfahrung der Coronakrise [zu] lernen und sie für einen Innovationsschub im Gesundheitswesen [zu] nutzen“ (Storm, 2021) [76]. Von Experten wird betont, dass trotz der Digitalisierung und dem Einsatz von Künstlicher Intelligenz (KI) im Gesundheitsbereich die Verantwortung und Kontrolle bei den Ärzten liegen sollte und die digitalen Dienste nur als Unterstützung dienen [65].

KI wird oft als Oberbegriff für Algorithmen, Methoden und Systeme, die ein scheinbar intelligentes Verhalten von Computern imitieren, benutzt. Die Anfänge von KI basieren eher auf regelbasierter Programmlogik. Durch die fortschreitende Digitalisierung und den Einsatz von Technologie wird eine Datenflut erzeugt. Diese ist im Alltag kaum zu bewältigen, beispielsweise ist es fast unmöglich, im Gesundheitswesen eine Diagnose unter Berücksichtigung aller Daten zu erstellen. Dennoch besteht die Hoffnung, dass durch den Nutzen von Big Data auch die Patientenversorgung im Gesundheitswesen verbessert werden kann [22]. Aktuell wird KI in vielen Themenbereichen von der Entscheidungsunterstützung wie z.B der Mustererkennung bis zur Automatisierung von Prozessen wie z.B. der Sprach- und Bildverarbeitung oder autonomen Robotern eingesetzt [22]. In der Sprachverarbeitung, auch Natural-Language-Processing (NLP) genannt, sind Large Language Models (LLM) mit der Veröffentlichung des ChatGPTs von OpenAI Ende des Jahres 2022 in den Fokus der Öffentlichkeit gerückt [28]. Durch die mediale Aufmerksamkeit ist das Interesse der Forschung an LLMs allgemein sowie auch an LLMs im Gesundheitsbereich gestiegen. Experten beschäftigen sich mit verschiedenen Anwendungsfällen.

„In der Medizin kann KI Aufgaben übernehmen, die viele ungern machen: Arztbriefe vorschreiben, Aktendokumentation oder Datenauswertung empfinden wir insgesamt oft

als langweilig und anstrengend. Und es nimmt vor allem Zeit mit den Patienten“ [65] behauptet, Prof. Dr. Buyx in dem Interview „Manche Aufgaben kann KI besser als wir“, zu der Frage ob Medizin von KI profitieren kann. LLMs als Teilgebiet der KI könnten eingesetzt werden, um die Patientenhistorie in Form einer Vor-Anamnese vor einem Behandlungstermin aufzunehmen. In einem menschenähnlichen Gespräch soll mithilfe eines LLM eine strukturierte Zusammenfassung der Symptome und Beschwerden der Patienten erzeugt werden, in der zusätzlich mögliche Vorerkrankungen oder die Einnahme von Medikamenten aufgenommen werden. Diese Zusammenfassung kann von den Patienten zum Termin mitgebracht werden und das Arztgespräch unterstützen. Hierdurch würde die Aktendokumentation vereinfacht. Vor allem aber entlastet sie das Arztgespräch und kann die Qualität der Diagnose verbessern, da bei der Vor-Anamnese menschliche Fehler wie das Überhören von Symptomen und Beschwerden verhindert werden. Dem Einsatz der KI sind immer noch Grenzen gesetzt. So kann hiermit kein Arztgespräch ersetzt werden, denn die Symptome sind zu vielfältig, als dass sie verlässlich zur Diagnose zusammengefasst werden können. Damit bleibt auch die Entscheidung über die Konsequenzen der Diagnose in der Verantwortung des Mediziners. Die KI erteilt keinen medizinischen Rat.

Um Bewerten zu können ob LLM für die Vor-Anamnese eingesetzt werden können, untersucht diese vorliegende explorative Studie den Auswahlprozess. Dabei werden ausgehend von einer Übersicht über mögliche freie LLMs, die eingesetzt werden könnten, Bewertungskriterien entwickelt. Außerdem wird untersucht mithilfe welcher Benchmarks die Bewertungskriterien evaluiert werden können, damit eine fundierte Entscheidung getroffen werden kann.

2 Stand der Technik

Da LLMs ein recht neues und unbekanntes Teilgebiet der KI sind, werden hier erstmal die Grundlagen betrachtet. Hierfür wird der Stand der Technik in die zwei Kategorien, die Definition und Einsatzmöglichkeiten der LLMs und die LLMs im medizinischen Bereich unterteilt.

2.1 Definition und Einsatzmöglichkeiten von LLMs

LLMs sind KI-Modelle, die speziell für die Verarbeitung von Sprache und Text entwickelt wurden. In der Sprachverarbeitung werden sie eingesetzt, um Texte zu verstehen, zu generieren und zu verarbeiten. Dabei werden komplexe neuronale Netzwerke eingesetzt, die Deep-Learning-Algorithmen und den Selbstaufmerksamkeitmechanismus nutzen. Deep-Learning ist der Teil des maschinellen Lernens, der das menschliche Lernverhalten mithilfe künstlicher neuronaler Netze und großen Datenmengen imitiert [92]. Das Training von LLMs erfolgt in der Regel anhand umfangreicher, nicht markierter Datensätze. Durch dieses Training entwickeln sie ein „Verständnis“ für den Kontext und die Struktur von Texten. Das „Verständnis“ der LLMs beruht auf der Fähigkeit, die menschliche Kommunikation anhand von stochastischen Berechnungen nachzuahmen. Mit diesem erworbenen „Verständnis“ können LLMs Muster im Text vorhersagen, dies ermöglicht ihre Anwendung in verschiedenen Aufgaben der Sprachverarbeitung. Sie können beispielsweise Fragen beantworten, Texte übersetzen oder Texte generieren. [49]

Insbesondere kann man LLMs in der Sprachverarbeitung für das Zusammenfassen von Texten einsetzten. Die Fachliteratur unterscheidet zwischen dem abstrakten und dem extraktiven Zusammenfassen. Während die KI beim abstrakten Zusammenfassen ihr Verständnis für menschenähnliche Sprache nutzt, um aus den zentralen Informationen einen neuen Text zu kreieren, in dem unter anderem Phrasen und Sätze aus dem originalen Text enthalten sind, filtert die KI beim extraktiven Zusammenfassen die Schlüsselsätze

und -phrasen des Originals und erstellt aus diesen einen neuen Text. In beiden Zusammenfassungen sind am Ende die wichtigsten Aussagen enthalten. [10]

Es gibt eine Vielzahl verschiedener LLMs auf dem Markt, die für verschiedene Anwendungsfälle geeignet sind. Zu den bekanntesten gehören PaLM von Google, LLaMA von Meta, GPT-4 von OpenAI und Stable LM von Stability AI. Dabei ist nicht jedes LLM auf einen einzigen Themenschwerpunkt beschränkt. GPT-3,5 von OpenAI findet beispielsweise vielfältige Anwendungen in unterschiedlichen Fachgebieten. Ein prominenter Einsatzbereich von GPT-3,5 ist beispielsweise der Chatbot ChatGPT. Darüber hinaus wird GPT-3,5 auch in den Bereichen Rechtswissenschaften, kreatives Arbeiten und Wissenschaft eingesetzt. Der wissenschaftliche Einsatz kann weiter unterteilt werden, wobei GPT-3,5 in den Disziplinen Informatik, Robotik und Medizin Anwendung findet. Im medizinischen Kontext kann GPT-3,5 Fragen beantworten und Informationen extrahieren. [44]

Zu Beginn kann ein LLM für verschiedene Themenbereiche eingesetzt werden. Um Sie für spezielle Anwendungsbereiche nutzbar zu machen, müssen Sie mit ausgewählten Datensätzen trainiert werden. Im Verlauf des Trainings werden Wahrscheinlichkeiten von Wortfolgen ermittelt und die Parameter des neuronalen Netzes werden entsprechend angepasst. In diesen Parametern ist jeweils ein Faktor gespeichert. Das neuronale Netz fasst alle Parameter zusammen und legt so die Entscheidungsregeln des LLM fest. Je mehr Parametern ein LLM hat, desto komplexere Muster können erkannt und mehr Informationen gespeichert werden. Typischerweise verfügt ein LLM über Billionen von Parametern [20]. Ein Teil des Trainings betrifft den Selbstaufmerksamkeitsmechanismus, der es ermöglicht, Teile der Eingabeinformation, sei es ein Wort oder ein Token, unabhängig von ihrer Position mit anderen Teilen der Eingangsinformation in Beziehung zu setzen. Auf diese Weise wird ein umfassenderes Verständnis für den gesamten Eingabetext erreicht [86] [49]. Das Training erfordert allerdings erhebliche Rechen- und Speicherressourcen, wobei dieser Bedarf mit der Anzahl der Parameter und der Datenmenge steigt. [49]

Um zu verhindern, dass für jede spezifische Aufgabe innerhalb eines Anwendungsbereiches ein neues LLM von Grund auf trainiert werden muss, wird die Methode der Feinabstimmung, auch finetuning, angewandt. Bei der Feinabstimmung wird ein bereits allgemein trainiertes LLM durch einen überwachten Lernvorgang mit markierten Daten speziell für die gewünschte Aufgabe trainiert. Während dieses Prozesses werden die Parameter des LLM gezielt optimiert, um die Leistungsfähigkeit für die spezifische Aufgabe zu verbessern. [49] [20] Ein Beispiel hierfür ist HuaTuo, ein LLM, das von einem Forschungs-

team in China feinabgestimmt wurde. Es basiert auf Metas LLaMA, die Feinabstimmung wurde unter Verwendung von unstrukturiertem und strukturiertem medizinischem Wissen aus dem „Chinesischen medizinischen Wissensgraph“ (CMeKG) durchgeführt. [87]

Trotz der vielfältigen Anwendungsmöglichkeiten von LLMs ist zu beachten, dass jeder Anwendungsbereich seine eigenen Herausforderungen mit sich bringt. Im medizinischen Bereich stellen die größten Herausforderungen, neben ethischen Fragen, Halluzinationen und Verzerrung der Modelle dar, auch als Bias bezeichnet. Halluzinationen treten auf, wenn LLMs fehlerhafte Informationen liefern, die aufgrund des flüssigen Textes schwer zu identifizieren sind [44]. Ein Bias kann durch fehlerhafte Daten entstehen [34]. So hat ein Forscherteam eine KI trainiert, die auf Bildern gesunde Haut von Hautkrebs unterscheiden soll. Da auf den Bildern mit dem Hautkrebs jedes Mal ein Lineal abgebildet war, hat die KI lediglich gelernt, ein Lineal mit Hautkrebs zu verknüpfen [73].

Eine weitere Unterscheidung zwischen LLMs liegt in der Lizenzierung. Es gibt open- und closed-source LLMs [60]. Zusätzlich können LLMs zur sogenannten freien Software gehören. Bei closed-source LLMs wird kein Einblick in den Quelltext gewährt [21]. Obwohl freie Software und Open-Source-Software ähnlich sind, dürfen sie nicht gleichgestellt werden [79]. Der Begriff Open Source bezieht sich auf Prinzipien, während es bei freier Software um vier wesentliche Freiheiten geht. Die Definition von freier Software besagt, dass „Nutzer die Freiheit haben, Software auszuführen, zu kopieren, zu verbreiten, zu untersuchen, zu ändern und zu verbessern“ [13]. Bei den Kriterien für Open Source geht es um die Lizenzierung des Quellcodes [79]. Unabhängig davon kann Software sowohl für die kommerzielle Nutzung als auch für die nichtkommerzielle Nutzung freigegeben sein.

2.2 LLMs im medizinischen Bereich

Das Europäische Medizinproduktgesetz teilt Produkte, die im medizinischen Bereich eingesetzt werden, in sogenannten Medizinprodukte und nicht Medizinprodukte ein. Wenn es sich um ein Medizinprodukt handelt, wird es einer der vier verschiedenen Risikoklassen I, IIa, IIb, und III zugeordnet. Ob ein Produkt oder eine Software als Medizinprodukt klassifiziert wird, hängt von der Zweckbestimmung ab. Handelt es sich um die reine Wissensbereitstellung wie beispielsweise ein Buch, ist dies kein Medizinprodukt. Werden aber Daten gesammelt und diese Daten beeinflussen die Diagnose oder Behandlung, handelt es sich um ein Medizinprodukt. Die Einordnung der Medizinprodukte in die Risikoklassen

erfolgt nach potenziellem Schaden, den sie bei Fehlern anrichten. Wobei Risikoklasse I dem geringsten Risiko entspricht und Klasse III dem größten. [55]

Im medizinischen Bereich werden mit Glass Health von Glass AI [39] und Ada von Ada Health [4] bereits zwei KI-Technologien eingesetzt. Glass Health fungiert als klinische Entscheidungshilfe, die Differentialdiagnosen und klinische Pläne erstellt. Zugleich ermöglicht sie den Zugriff auf eine gemeinschaftliche Bibliothek mit geteiltem medizinischem Wissen. Glass Health richtet sich ausschließlich an medizinisches Personal und kombiniert ein LLM mit evidenzbasierten klinischen Richtlinien, die von Ärzten erstellt und gepflegt werden. Glass Health ist ein in Amerika entwickltes Produkt und unterliegt somit nicht dem Europäischen Medizinproduktgesetz. Ada, als Medizinprodukt der Klasse I und II, steht hingegen jedem zur Verfügung. Die genaue KI-Technologie hinter Ada ist nicht veröffentlicht, jedoch handelt es sich wahrscheinlich um eine regelbasierte Programmlogik. Mit Ada können Diagnosen auf Grundlage von angegebenen und abgefragten Symptomen und Beschwerden erstellt werden. Neben den Diagnosen werden auch die Symptome und Beschwerden gespeichert, die zu der jeweiligen Diagnose geführt haben, um sie bei einem Arztgespräch vorzulegen.

Ein weiteres bekanntes LLM im medizinischen Bereich ist Med-PaLM 2. Dieses Modell basiert auf PaLM 2 von Google und wurde mit Hilfe von medizinischem und klinischem Wissen trainiert. Med-PaLM 2 kann Fragen beantworten und Erkenntnisse aus verschiedenen medizinischen Texten zusammenfassen. Es ist auch das erste Sprachmodell, das bei Fragen im Stil des U.S. Medical Licensing Exam auf „Experten“-Niveau abschneidet. Das nächste Ziel der Entwickler ist es Med-PaLM 2 multimodale Fähigkeiten hinzuzufügen, um Informationen wie Röntgenbilder und Mammogramme zu synthetisieren und bestenfalls die Ergebnisse für Patienten zu verbessern. [36] [75]

Es gibt weitere LLMs, die für den wissenschaftlichen und medizinischen Bereich feinabgestimmt oder trainiert wurden. Diese wurden im Zusammenhang mit Studien entwickelt und werden im nächsten Kapitel behandelt.

3 Stand der Wissenschaft

Durch die zunehmende Forschung im Bereich der LLMs muss der Stand der Wissenschaft betrachtet werden, um zu analysieren welche Studien bereits durchgeführt wurden. Hierfür kann dieser in drei Bereiche geteilt werden. Der erste Bereich sind die Studien die sich mit den LLMs im Allgemeinen beschäftigt haben. Im zweiten Abschnitt werden dann die Studien zu LLMs im medizinischen Bereich betrachtet und im letzten die Studien, die die Bewertung der LLMs untersucht haben.

Für die Bewertung von LLMs werden häufig sogenannte Benchmarks herangezogen. Benchmarks sind standardisierte Leistungstest, die meist auf spezifischen Sprachverarbeitungsaufgaben basieren und verschiedene Test Datensätze beinhalten. Mithilfe der Testdaten kann die Leistung der Software, in diesem Falle des LLM gemessen werden. Somit können zwei LLMs miteinander verglichen werden. Weiterhin erhält man quantitative Daten zur Beurteilung der Leistungsfähigkeit der getesteten Modelle. [68]

3.1 Studien über LLMs

Die Studie „A Comprehensive Overview of Large Language Models“ [60] bietet einen umfassenden Einblick in die neuesten Entwicklungen im Forschungsbereich der LLMs. Sie stellen eine systematische Übersicht bereit, die als schnelles und umfassendes Nachschlagewerk für Forscher dienen soll. Die Einblicke in die aktuelle Forschung sind als Zusammenfassungen bestehender Arbeiten konzipiert und sollen die Forschung im Bereich der LLMs unterstützen. Die Forschungsbeiträge, die in der Studie behandelt werden, erstrecken sich über verschiedene Themen, darunter architektonische Innovationen der zugrundeliegenden neuronalen Netze sowie Benchmarks, Trainingsdaten und verschiedene LLMs. [60]

Eine Studie, die sich mit der Textzusammenfassung befasst, trägt den Titel „Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct,

Falcon-7b-instruct, and OpenAI Chat-GPT Models“ [10]. In dieser Untersuchung wurden verschiedene LLMs hinsichtlich ihrer Fähigkeiten zur Textzusammenfassung analysiert. Zur Bewertung wurden gängige Benchmarks wie der Bilingual Evaluation Understudy (BLUE) Score, der Recall-Oriented Understudy for Gisting Evaluation (ROUGE) Score und der Bidirectional Encoder Representation from Transformers (BERT) Score verwendet. Die Ergebnisse auf Basis der Benchmarks zeigen, dass das LLM text-davinci-003 die besten Leistungen erzielte. Es wurde die Annahme getroffen, dass die Modellarchitektur, die Menge der Trainingsdaten und die Anzahl der Parameter eine entscheidende Rolle für die Qualität der Textzusammenfassung spielen, da text-davinci-003 in der Studie als das Sprachmodell mit den meisten Parametern und der größten Menge an verwendeten Textdaten herausragte. [10]

3.2 Studien über LLMs im medizinischen Bereich

Ein Einsatzgebiet von LLMs ist der medizinische Bereich, hier existieren die verschiedensten Studien. Allerdings gibt es keine Studien zum Einsatz von LLMs im Bereich der Vor-Anamnese.

Da ChatGPT das LLM mit der größten medialen Aufmerksamkeit und der größten Anwendergruppe ist, haben sich einige Studien speziell mit diesem Sprachmodell befasst. Darunter sind zwei, die den Nutzen von ChatGPT im medizinischen Bereich untersucht haben. Die erste Studie mit dem Titel „Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models“ [45] befasst sich mit der Anwendung von ChatGPT am United States Medical Licensing Exam (USMLE). Die Ergebnisse zeigten, dass ChatGPT es schafft das USMLE mit 60% zu bestehen. Die Forscher sind der Meinung, dass LLMs das Potenzial haben, Lernende im medizinischen Bereich zu unterstützen.

In der zweiten Studie mit dem Titel „Future of the Language Models in healthcare: the role of ChatGPT“ [85] wird untersucht, inwieweit ChatGPT das Gesundheitswesen unterstützen kann. Auch hier sind sich die Forscher einig, dass ChatGPT und LLMs im Allgemeinen das Potenzial haben, im Gesundheitswesen eingesetzt zu werden. Allerdings betonen die Forscher die Notwendigkeit, die Grenzen der Technologie wie Halluzinationen, Bias und die Schwierigkeiten bei der korrekten Einordnung von Informationen, im Kontext zu berücksichtigen. Zudem wird betont, dass die Technologie eher als Unterstützung für das Gesundheitspersonal gesehen wird.

Ein weiterer wichtiger Beitrag ist „Large language models in medicine: the potentials and pitfalls “ [63], dieser soll die Anwender aus dem Gesundheitsbereich unterstützen. Daher wird veranschaulicht wie die Entwicklung aussieht, aber auch welche Schwierigkeiten und Grenzen der Einsatz von LLMs im Gesundheitsbereich mit sich bringen. Außerdem wird ein Überblick über die aktuellen und die Applikationen mit Potenzial gegeben. Auch werden verschiedene LLMs vorgestellt, die für den medizinischen Bereich eingesetzt werden. Zu diesen Modellen gehören BioGPT und BioMedLM die auf GPT von OpenAI basieren. Vier kleinere Modelle die auf BERT basieren, sind PubMedBERT, BioBERT, BioLinkBERT und ClinicalBERT. Zwei weitere Modelle mit einer großen Anzahl an Parametern, über 540 Billionen, sind Flan-PaLM und das bereits erwähnte Med-PaLM, beide basieren auf Googles PaLM. Außerdem werden PMC-LLaMA, welches auf LLaMA basiert und GatorTron vorgestellt. GatorTron ist ein LLM, welches Patienteninformationen aus unstrukturierten elektronischen Gesundheitsdaten extrahieren kann [95]. Neben der Vorstellung der LLMs werden auch die Schwierigkeiten, die mit dem Einsatz von LLMs einhergehen wie der Bias betrachtet. Aber auch die fehlende HIPPA-Konformität ist ein Problem. HIPPA ist ein amerikanisches Gesetz, welches einen Standard, zum Schutz von sensiblen Gesundheitsdaten und zum Schutz vor Weitergabe ohne die Einwilligung und das Wissen des Patienten etabliert [67]. Somit können keine geschützten Patientendaten geteilt werden. Eine weitere Schwierigkeit, die den Einsatz der LLMs für die medizinische Bildung betrifft, ist die nicht ausreichende Personalisierung. Denn die LLMs sind mit bereits existierenden Arbeiten trainiert worden. Auch wurden bereits zahlreiche Diskussionen über den Zugang und den Einsatz von LLMs im medizinischen Bereich geführt. Die wichtigste Erkenntnis ist, dass trotz der vielen Einsatzmöglichkeiten wie in der klinischen Administration, in der Forschung oder der Bildung, viele Schwierigkeiten und Herausforderungen mit LLMs weiterbestehen bleiben. Zu den Schwierigkeiten und Herausforderungen gehören der Bias, die Datenqualität, die unvohersehbaren Ausgaben, der Schutz der Privatsphäre des Patienten und die ethischen Bedenken.

Die Studie „Large language models in medicine“ [83] richtet sich ebenfalls an Anwender aus dem medizinischen Bereich. Die Studie soll als Leitfaden dienen, auf Grundlage dessen Entscheidungen getroffen werden können, ob und in welcher Weise LLM-Technologie im Gesundheitswesen eingesetzt werden kann, damit ein Nutzen für Ärzte und Patienten entsteht. Es werden die Stärken und Grenzen der LLMs aufgezeigt. Eine wichtige Erkenntnis dieser Studie ist, dass potentielle Risiken bei Experten und der Gesellschaft große Bedenken gegenüber der Sicherheit, der Ethik und dem potentiellen Ersatzes von Menschen hervorruft. Forscher betonen, dass der autonome Einsatz von LLMs derzeit

nicht möglich ist, Aber validierte Anwendungen als wertvolle Unterstützung zur Verbesserung des Gesundheitswesens gesehen werden. Zu einer erfolgreichen Validierung gehören pragmatische klinische Studien, die einen echten Nutzen bei minimaler Verzerrung und transparenter Berichterstattung nachweisen.

Neben den Studien, deren Ergebnis eine Übersicht ist, existieren im medizinische Bereich auch Studien, in denen LLMs entwickelt oder feinabgestimmt werden. Zu diesen Studien gehört „BioBERT: a pre-trained biomedical language representation model for biomedical text mining“ [46], in welcher das bereits erwähnte LLM BioBERT trainiert wurde. BioBERT wurde auf einem großen biomedizinischem Textkorpus vortrainiert und basiert auf BERT. BioBERT wurde entwickelt, da durch das rapide Wachstum an biomedizinischen Dokumenten das biomedizinische Text Mining an Wichtigkeit gewinnt. Biomedizinisches Text Mining ist eine Sammlung von Algorithmen, die es ermöglichen unstrukturierte Texte zu verarbeiten und die wichtigsten Zusammenhänge zu erkennen und somit die Textverarbeitung zu beschleunigen [93]. Für die Validierung von BioBERT wurde das Modell mit drei repräsentativen biomedizinischen Text Mining Aufgaben mit verschiedenen Datensätzen getestet. Die Text Mining Aufgaben gehören zu den folgenden Bereichen: Der erste ist die Entdeckung von benannten Entitäten (NER), der zweite die Extraktion von Beziehungen und als letztes der Bereich Antworten auf Fragen, siehe Anhang A.3. Die Tests zeigen, dass BioBERT im medizinischen Bereich bessere Precision, Accuracy und Recall Wert erreicht hat als BERT und der State-of-the-Art. Wobei der State-of-the-Art das beste und modernste LLM zum Zeitpunkt der Tests ist [27].

Die Studie „SCIBERT: A Pretrained Language Model for Scientific Text“ [11] beschäftigt sich mit der Entwicklung des LLM SciBERT. SciBERT ist ebenfalls ein vortrainiertes Modell, welches auf BERT basiert. Für das Training wurde ein unüberwachter Ansatz und für die Trainingsdaten eine Textsammlung aus verschiedenen wissenschaftlichen Veröffentlichungen, die zu verschiedenen Bereichen gehören, gewählt. Für die Validierung wurde SciBERT in verschiedenen Aufgaben der Sprachverarbeitung mit Datensätzen aus dem wissenschaftlichen Bereich getestet. Die fünf Aufgabenbereiche, mit denen getestet wurde, sind NER, Evidenzbasierte medizinische Informations Extraktion (PICO), Analyse von Abhängigkeiten, Relation Extraction und Textklassifikation, siehe Anhang A.3. Dabei schneidet SciBERT besser ab als BioBERT und das State-of-the-Art.

3.3 Studien zur Bewertung von LLMs

Die bereits existierenden LLMs müssen bewertet werden, damit das für die Anfertigung der Vor-Anamnese passendste ausgewählt werden kann. Eine Voraussetzung für die Auswahl ist, dass das zu bearbeitende Problem klar definiert und spezifiziert ist [12]. Im Kontext der Vor-Anamnese stellt sich das Problem wie folgt dar: mithilfe eines LLMs sollen Symptome und Beschwerden gesammelt und im Anschluss zusammengefasst werden. Um die Qualität der LLMs zu bewerten, eignen sich wie bereits beschrieben Benchmarks. In verschiedenen Studien werden unterschiedliche Benchmarks in der medizinischen und wissenschaftlichen Domäne vorgestellt.

Der erste Versuch, ein NLP Benchmark in der biomedizinischen Domäne zu erstellen, [37], erfolgt in folgender Studie „Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets“ [66]. Inspiriert durch den Erfolg der General Language Understanding Evaluation (GLUE), die für die Bewertung der LLMs im allgemeinen Sprachverständnis eingesetzt wird, wurde in der Studie das Biomedical Language Understanding Evaluation (BLUE)-Benchmark, für das biomedizinische Verständnis entwickelt. BLUE verwendet aus fünf verschiedene Aufgabenbereichen zehn Datensätzen, die aus verschiedenen medizinischen und klinischen Textsammlungen bestehen, siehe Anhang A.3. Die Datensätze wurden ausgewählt, da sie in der BioNLP Gemeinschaft viel benutzt werden. BLUE soll die Forschung im biomedizinischen Bereich und die Entwicklung von Sprachrepräsentationen erleichtern. In der Studie wurde ebenso die Entwicklung von verschiedenen BlueBERT-Modellen behandelt. Die BlueBERT LLMs wurden auf PubMed Abstrakten und klinischen Notizen aus dem Datensatz MIMIC-III trainiert. In der Evaluation haben diese BlueBERT-Modelle besser bewertet als das State-of-the-Art, ELMo und das zugrundeliegende BERT-Modell.

Eine weitere Studie, innerhalb welcher ein Benchmark entwickelt wurde, ist „Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing“ [37], in dieser wurde das Biomedical Language Understanding & Reasoning Benchmark (BLURB) veröffentlicht. Außerdem wurde ebenso das bereits erwähnte PubMedBERT trainiert und getestet. Wie auch BLUE besteht BLURB aus bekannten Datensätzen, die sechs Aufgabenbereiche mit dreizehn Datensätzen abdecken, siehe Anhang A.3. Auch in dieser Studie wurde PubMedBERT mit PubMed Abstrakten trainiert und hat besser abgeschnitten als BERT, RoBERTa, BioBERT, SciBERT, ClinicalBERT und BlueBERT. BLURB wurde entwickelt, um die Untersuchung der Annahme, dass das Domänspezifische Vortrainieren von generellen Modellen profitieren kann, zu unterstützen. Das Ergebnis der Studie ist,

dass Modelle, die von Grund auf für die biomedizinische Domäne trainiert wurden, besser abschneiden.

Mit den Grenzen der Evaluation von LLMs beschäftigt sich die Studie „Large language models encode clinical knowledge“[75]. In dieser wird das Benchmark MultiMedQA vorgestellt und das LLM Med-PaLM, welches von Google veröffentlicht wurde. MultiMedQA besteht aus sechs bestehenden Datensätzen für die Beantwortung von medizinischen Fragen in den Bereichen der medizinischen Prüfungen, wie dem USMLE, der Forschung und medizinischen Verbraucherfragen. Außerdem ist noch ein neuer Datensatz HealthSearchQA enthalten. Dieser umfasst medizinische Fragen, die online gestellt wurden. Mit dem Benchmark werden verschiedene auf PaLM basierende Modelle evaluiert. Flan-PaLM erreicht bei der Evaluierung State-of-the-art Genauigkeit, allerdings zeigt die menschliche Evaluation, dass noch große Unterschiede zwischen dem wissenschaftlichen Konsens und den Antworten des LLM bestehen. Daher wird PaLM noch weiter feinabgestimmt und daraus resultiert Med-PaLM. Med-PaLM schneidet bei der Beantwortung der Fragen besser ab und erreicht laut der menschlichen Evaluation, dass 92,6% der Antworten mit dem wissenschaftlichen Konsens übereinstimmen. Auch wenn Med-PaLM eine gute Quote erfüllt, stimmen bei der Beantwortung von Experten 92,9% der Antworten mit dem wissenschaftlichen Konsens überein. Die Forscher in dieser Studie sind sich einig, dass die Evaluation durch den Menschen die Grenzen der heutigen Modelle aufzeigt und die Bedeutung von Bewertungsmaßnahmen und -funktionen unterstreicht.

Die verschiedenen Aufgabenbereiche, die bei den Evaluationen der LLMs im medizinischen Bereich und den Benchmarks betrachtet werden, sind die folgenden.

3.3.1 Entdeckung von benannten Entitäten (NER)

Im Aufgabenbereich NER sollen Entitäten, die von Interesse sind, erkannt und vorhergesagt werden [37]. Eine Entität kann ein Wort oder auch eine Reihe von Wörtern sein, die immer die selbe Bedeutung haben. NER wird verwendet um das Thema eines Textes zu erkennen. Im Gesundheitswesen kann NER eingesetzt werden um die Versorgungsstandards der Patienten zu verbessern und die Arbeitslast der im Gesundheitswesen Tätigen zu reduzieren, indem beispielsweise wichtige Informationen aus Laborberichten extrahiert werden können [54]. Für die Evaluation wird die strikte Version von Precision, Recall und dem F1-Score verwendet [66].

3.3.2 Evidenzbasierte medizinische Informations Extraktion (PICO)

Als PICO wird ein Schema in der evidenzbasierten Medizin bezeichnet, mit dem Antworten auf konkrete therapeutische Fragen gefunden werden sollen. PICO als Akronym steht für die Elemente Patient (P), Behandlung (I), Alternativmaßnahme (C) und Behandlungsziel (O). Mithilfe des PICO-Schemas wird versucht, die Antwort auf eine Fragestellung möglichst übersichtlich darzustellen [42]. In dem Aufgabenbereich wird getestet, wie gut ein Modell die PICO-Elemente innerhalb einer Zusammenfassung, die ein klinisches Vorgehen beschreibt, bestimmen kann. Dabei kann eine Entität auch zu mehreren PICO-Elementen gehören. [37]

3.3.3 Extraktion von Beziehungen

Für die Extraktion von Beziehungen soll zwischen Entitäten in einem Satz die semantische Beziehung vorhergesagt werden. Die Beziehung zwischen medizinischen Begriffen ist wichtig für die Identifizierung von medizinischem Wissen. Auch kann die Erkennung von Beziehungen zwischen medizinischen Krankheiten, Tests und Behandlungen, die Qualität der Patientenversorgung verbessern [64]. Die Validierung findet durch die Standard micro-average Precision, Recall und F1-Score Metriken statt. [37]

3.3.4 Satzähnlichkeit

Für die Überprüfung der Satzähnlichkeit sollen Satzpaare verglichen [37] und somit die Übereinstimmung zweier Texte überprüft werden [32]. Für die Validierung wird der Pearson correlation Coefficients verwendet.

3.3.5 Dokumentenklassifikation

Bei der Dokumentenklassifikation soll für Dokumente, welche ein Text, Foto oder auch ein eingescanntes Dokument sein können, ein oder mehrere Klassen vorhergesagt werden. Dadurch sollen die Dokumente einfacher zu suchen, filtern und analysieren sein. Um den Dokumentenfluss zu erleichtern werden im Zuge der Digitalisierung im Gesundheitsbereich den eingescannten Dokumenten mithilfe verschiedener Dokumentenklassifizierungen Klassen zugeordnet [71]. So soll beispielsweise im Datensatz Hoc den Dokumenten die zehn bekannten Kennzeichnungen und Merkmale von Krebs zugeordnet werden. [66]

3.3.6 Antworten auf Fragen

In dem Aufgabenbereich der Beantwortung von Fragen soll mithilfe eines Textabschnittes eine Frage mit ja, nein bzw. ja, nein, vielleicht beantwortet werden. [37]

3.3.7 Analyse von Abhängigkeiten

Die Analyse von Abhängigkeiten wird verwendet, um die grammatikalische Struktur eines Textes zu bestimmen. Hierfür wird die Abhängigkeit verschiedener Phrasen eines Texten untersucht. [74]

3.3.8 Textklassifikation

In der Textklassifikation soll, ähnlich der Dokumentenklassifikation, für ein Textdokument, einen Satz oder eine Phrase eine vordefinierte Klasse vorgeschlagen werden. Der Unterschied zu der Dokumentenklassifikation ist, dass hier kleinere Textabschnitte verwendet werden und auch nur Textdokumente. [61]

3.3.9 Schlussfolgerungsaufgabe

Für die Schlussfolgerungsaufgabe werden Satzpaare vorgegeben, von denen einer die Prämisse ist und der andere die Hypothese. Für die Satzpaare soll vorhergesagt werden, ob die Prämisse, die Hypothese einschließt, der Hypothese widerspricht oder keins von beidem. Zum Evaluieren wird wird die standard Accuracy verwendet. [66]

4 Vorauswahl geeigneter LLMs

LLMs können nach verschiedenen Kriterien bewertet werden. Allgemeine Kriterien sind die Lizenzbedingungen, der Datenschutz, aber auch die Trainingsdaten und der Anwendungsfall, für welchen sie trainiert oder feinabgestimmt wurden.

In dieser explorativen Studie sollen freie LLMs untersucht werden, da neben den günstigeren Kosten die Freiheiten gegeben sind, dass die LLMs ausgeführt, untersucht und geändert werden dürfen. Ein weiterer Punkt, der für die Beschränkung auf freie LLMs spricht, ist, dass der Quellcode von nicht freien Modellen eher abgeschottet ist und keine Einblicke gewährt. Somit müssen die LLMs ermöglichen, dass sie ausgeführt, untersucht, geändert, kopiert und verbessert werden dürfen. Die einzige Freiheit, die nicht erfüllt werden muss, ist die der Weiterverbreitung, da es sich bei dieser Arbeit um eine explorative Studie handelt. Auch das Recht auf kommerzielle Nutzung ist nicht essentiell, wäre aber von Vorteil, falls in einer späteren Arbeit tatsächlich ein LLM für die Vor-Anamnese im Gesundheitsbereich entwickelt werden würde.

Da die explorative Studie den Gesundheitsbereich betrifft, ist ein wichtiges Kriterium bei der Auswahl auch der Datenschutz. Bei dem Thema Datenschutz wird betrachtet, wie die Daten verarbeitet oder weitergegeben werden. Da im medizinischen Bereich sensible personenbezogene Daten verarbeitet werden, ist es wichtig, dass bei einem Einsatz eines LLMs im Gesundheitsbereich gewährleistet ist, dass das Grundrecht an den eigenen Daten nicht verletzt wird. Dies kann auch für die Trainingsdaten gelten, je nachdem wie diese aufbereitet wurden. Bei einer kompletten Anonymisierung der Trainingsdaten ist die Weiterverarbeitung datenschutzrechlich kein Problem. Allerdings ist der Einsatz anonymisierter Daten nicht von Vorteil, da bei falsch annotierten Datensätzen eventuell fehlerhafte Krankheitsbilder antrainiert werden können. Und durch den fehlenden Zugriff auf die Originaldaten die Datenqualität schwer gewahrt werden kann. Bei der Auswahl der LLMs gestaltet sich die Betrachtung der Trainingsdaten als eher schwierig, da ihnen nicht angesehen werden kann, wie sie verarbeitet wurden. Daher wird dies bei der Auswahl nicht beachtet. Wie bereits erläutert muss bei dem potenziellen Einsatz von LLMs

im Gesundheitsbereich sichergestellt werden, dass das Grundrecht an der persönlichen Daten nicht verletzt wird. [17]

Bei LLMs, die nicht für die allgemeine Domäne trainiert wurden, ist wichtig, dass die Feinabstimmung oder das Training für den speziellen oder einen ähnlichen Anwendungsfall stattgefunden hat. In dieser Arbeit sollten die LLMs entweder für den Einsatz im medizinischen Bereich oder für die Textzusammenfassung trainiert worden sein. Auch LLMs, die für mehr Sprachverarbeitungsaufgaben als nur die Textzusammenfassung trainiert wurden oder für den Einsatz in der Wissenschaft, können für den Einsatz im Gesundheitsbereich in Betracht gezogen werden.

Um einen Überblick über alle zur Verfügung stehenden LLMs zu erhalten, wurde aus den LLMs, die in den vorherigen Kapiteln vorgestellt wurden, die folgende Übersicht erstellt, siehe Anhang A.1. Im ersten Schritt wurden alle LLMs, die nicht der Definition der freier Software entsprechen, ausgeschlossen. Bei zwei LLMs und deren fein-abgestimmten Versionen, gab es laut den Quellen Unstimmigkeiten bezüglich der Lizenz. Somit wurden Bloom/Bloom-Z und Claude/Claude 2 nochmal überprüft, ob sie der Definition von freien LLMs entsprechen und die Freiheiten erfüllen. Bloom-Z ist eine feinabgestimmte Version von Bloom [14] und beide Modelle dürfen unter der BigScience Lizenz verwendet werden. Die BigScience Lizenz erlaubt die Freiheiten der freien Software mit der Einschränkung, dass kein klinischer Rat gegeben oder die Symptome interpretiert werden dürfen, siehe Use Restriction (l) [3]. Bei den Modellen Claude und Claude-2, sieht das anders aus. Beide sind im deutschen Raum nicht zugänglich und können somit in dieser Studie nicht berücksichtigt werden [8].

Die restlichen LLMs (siehe Anhang A.2) wurden auf ihren trainierten Anwendungsfall und die Feinabstimmung überprüft. Die ersten Modelle, die aufgrund der Feinabstimmung mit chinesischen Daten verworfen werden, sind XuanYuan 2.0, PanGu-α, CPM-2 und das bereits erwähnte HuaTuo. Alle vier Modelle sind, mit chinesischen Daten trainiert worden. XuanYuan 2.0 ist eine feinabgestimmte Version von Bloom und wird im chinesischen Finanzmarkt als Grundlage für einen Chatbot verwendet. HuaTuo wurde zwar für den medizinischen Bereich feinabgestimmt, ist auf Grund der Sprachbarriere nicht geeignet.

Ein weiterer Anwendungsfall, für den viele LLMs trainiert und feinabgestimmt wurden, ist die Codeintelligenz. Die LLMs wurden unter anderem mit Programmiersprachen trainiert und werden eingesetzt um Code zu generieren und zu interpretieren. Da das

interpretieren und generieren von Code kein Anwendungsfall ist, der dem Zusammenfassen von Texten ähnelt oder Anwendung im medizinischen Bereich findet, kommen die folgenden LLMs: CodeGen, Code LLaMA, Wizard Coder, CodeT5+, StarCoder, Phi-1 und StableLM nicht in Frage. Ein weiteres Modell das zwar nicht von Anfang an für die Codeintelligenz trainiert wurde, aber auch aus diesem Grund verworfen wird, ist StableLM.

Auch die LLMs die für die multilinguale Domäne trainiert wurden, sind weder für den Gesundheitsbereich noch für die Textzusammenfassung tauglich. LLMs, die in diesem Bereich feinabgestimmt oder trainiert wurden, sind gut in der Textgeneration und Textübersetzung von verschiedenen Sprachen. Zu diesen ungeeigneten Modellen gehören MT0, mT5, Guanaco-65B und Bloom-Z.

Zwei weitere LLMs die aufgrund ihrer Feinabstimmung von LLaMA nicht geeignet sind, sind Orca und Goat. Orca wurde für die Argumentation feinabgestimmt, somit kann Orca gut eingesetzt werden, wenn Entscheidungen getroffen oder Probleme gelöst werden sollen. Goat wurde für die Berechnung von verschiednen Rechenaufgaben feinabgestimmt.

Ein LLM, das prinzipiell für die Anwendung im wissenschaftlichen Bereich eingesetzt werden kann, ist Galactica. Galactica wurde für den wissenschaftlichen Einsatz trainiert und kann argumentieren, mathematische Aufgaben lösen sowie medizinische Fragen beantworten. Somit wäre es eigentlich gut geeignet, da es schon in der medizinischen Domäne trainiert wurde. Allerdings hat Meta Galactica abgeschaltet, da es einen Hang zu Halluzinationen hat. [16]

Die folgende Übersicht zeigt die LLMs, die den definierten Kriterien entsprechen, um in der medizinischen Domäne und für die Textzusammenfassung eingesetzt zu werden.

LLM	vortrainiert	feinabgestimmt	Besonderheiten
T5 [33]	ja	nein	
T0 [70]	ja	nein	basiert auf T5
OPT-IML [43]	ja	ja	fein-abgestimmte Version von OPT
GLM [29]	ja	nein	pre-trained mit autoregressive blank filling objective
OPT [98]	ja	nein	evtl. kein Access
UL2 [31]	ja	ja	ähnlich zu T5, gleicher Satz tokenizer verschiedene Objektive und C4 Korpus

LLM	vortrainiert	feinabgestimmt	Besonderheiten
Tk-Instruct [89]	ja	ja	build upon T5 models
GPT-NeoX-20B [15]	ja	nein	architecture ähnelt GPT-3
BLOOM [59]	ja	nein	Architektur ähnelt GPT-3 Architektur
LLaMA 2 [57]	ja	ja	für Chat-Anwendung fein-abgestimmt
MPT [82]	ja	nein	es existieren verschiedene fein-abgestimmte Versionen von MPT
Koala [47]	ja	ja	fein-abgestimmter Chat-Bot basierend auf LLaMA
WizardLM [94]	ja	ja	fein-abgestimmtes LLaMA Chatbot
Vicuna [101]	ja	ja	fein-abgestimmter Chat-Bot basierend LLaMA
Alpaca [80]	ja	ja	fein-abgestimmtes LLaMA
LLaMA [84]	ja	nein	
Falcon [6]	ja	nein	keine
30B-Lazarus [18]	ja	nein	basiert auf LLaMA
BERT [41]	ja	nein	es gibt viele vortrainierten Modelle die verwendet werden können
FLAN-T5 [30]	ja	ja	von T5 trained on Flan-Collection
Dolly [25]	ja	ja	basiert auf pythia-12b
OpenChatKit [24]	ja	nein	Sammlung an LLMs
BioGPT [50]	ja	nein	
BioMedLM [63]	ja	nein	in Zusammenarbeit mit MosaicLM
BioBERT [46]	ja	ja	
BioLinkBERT [96]	ja	nein	

LLM	vortrainiert	feinabgestimmt	Besonderheiten
ClinicalBERT [7]	ja	nein	
PMC-LLaMA [63]	ja	ja	
GatorTron [95]	ja	nein	
SciBERT [11]	ja	nein	

Tabelle 4.1: Überblick über verwendbare LLMs

5 Bewertung der LLMs

Für die Evaluation, welches der freien LLMs aus der Übersicht in Tabelle 4.1 am besten für die Erstellung einer Zusammenfassung aus einer Vor-Anamnesegespräch geeignet ist, werden weitere Bewertungskriterien benötigt. Da der Stand der Forschung hier eine Lücke aufweist, müssen diese Kriterien formuliert werden. Um Bewerten zu können wie leistungsfähig ein LLM im Bereich eines Kriteriums ist, können die passenden Benchmarks verwendet werden. Mithilfe dieser Kriterien und den passenden Benchmarks soll eine fundierte Auswahl getroffen werden.

5.1 Kriterien zur Bewertung

Die Bewertung der LLMs soll auf Basis der folgenden Kriterien erfolgen: die Qualität des Vor-Anamnesegesprächs, das medizinische Verständnis des LLMs sowie das Erinnerungsvermögen des LLMs. Je besser das LLM in den genannten drei Kriterien abschneidet, desto besser wird die Qualität im vierten Kriterium, der Qualität der Zusammenfassung des Vor-Anamnesegespräches, ausfallen.

Für die Bewertung der Qualität des Vor-Anamnesegespräches durch das LLM soll darauf geachtet werden, dass das Gespräch einem Anamnesebogen ähnelt. Es sollte neben den Symptomen und Beschwerden auch nach möglichen Vorerkrankungen oder der Einnahme von Medikamenten gefragt werden. Außerdem sollte eine Rückfrage kommen, wenn noch weitere Informationen benötigt werden. Wenn beispielsweise ein Patient angibt, dass er Bauchschmerzen hat, sollte eine Nachfrage erscheinen, wo genau die Bauchschmerzen sind, ob er die Bauchschmerzen noch genauer beschreiben kann und wie lange diese schon anhalten. Denn nur wenn die Vor-Anamnese möglichst umfassend und detailliert ist, kann die generierte Zusammenfassung den Arzt im Arztgespräch entlasten.

Auch wenn aus den Symptomen, Beschwerden und möglichen Vorerkrankungen „nur“ eine Zusammenfassung erstellt wird und kein medizinischer Rat gegeben oder eine medizinische Interpretation stattfinden soll, muss das LLM ein medizinisches Verständnis vorweisen, um aus den Inhalten alle wichtigen Informationen filtern zu können. Gibt ein Patient beispielsweise an, dass er an Nephromegalie leidet und dem LLM fehlt das Verständnis, dass es sich hierbei um eine Nierenvergrößerung handelt, kann es dazuführen, dass das LLM dies in der Zusammenfassung nicht erwähnt, weil diese Information nicht als wichtig angesehen wird. Das medizinische Verständnis wird des Weiteren noch benötigt, damit das LLM im Kontext der Vor-Anamnese wichtige Zusammenhänge und Implikationen richtig interpretieren kann und alles dafür relevante in der Zusammenfassung ausführt. Beispielsweise wäre wichtig, wenn der Patient angibt, dass er Bauchschmerzen hat und mit einer Nachfrage spezifiziert, dass die Bauchschmerzen im rechten Unterbauch, eher ein Druckschmerz sind und schon mehrere Stunden anhalten und zunehmen, dass dieser Zusammenhang auch in der Zusammenfassung dargestellt wird. Wenn der Patient dazu noch angegeben hat, dass sein Blinddarm operativ entfernt wurde, ist das eine weitere relevante Information im Zusammenhang mit Bauchschmerzen und somit unverzichtbarer Bestandteil der Zusammenfassung, da die Symptome unter anderem zu einer Blinddarmentzündung passen. So kann der Arzt eine Blinddarmentzündung ausschließen [35].

Ein Kriterium, das über den medizinischen Bereich hinausgeht, ist die Qualität der Zusammenfassung. Das LLM sollte auch außerhalb des medizinischen Bereiches in der Lage sein, Texte präzise und akkurat zusammenzufassen, damit alle wichtigen Informationen in der Zusammenfassung enthalten sind. Damit einher geht, dass das LLM genügend Erinnerungsvermögen mitbringen muss. Denn wenn das Vor-Anamnesegespräch nur zur Hälfte in der Zusammenfassung wieder gegeben wird, weil das LLM nach zu vielen Nachfragen die Informationen vom Anfang „vergessen“ hat, ist die Zusammenfassung für den Arzt keine Entlastung, vielmehr muss dieser mehr Zeit darauf verwenden, herauszufinden, welche Informationen fehlen.

5.2 Benchmarks

Die für die Textzusammenfassung gängigsten Benchmarks sind BLEU, BERT und ROUGE [10]. Die beiden Benchmarks BLEU und ROUGE können die Zusammenfassungsfähigkeit eines LLMs sprachunabhängig bewerten, betrachten dabei aber nicht die Se-

mantik. Bei BLEU haben beispielsweise die Artikel die gleiche Relevanz in einem Satz wie das Subjekt und das Verb [9]. Eine weitere Schwachstelle von ROUGE ist, dass die Bewertung der Qualität der Zusammenfassung von der Länge der Zusammenfassung beeinflusst wird [56]. Das Benchmark BERT basiert auf dem Modell BERT und hat zwar ein besseres Verständnis für den Kontext des Textes, kann aber nur für wenige Sprachen eingesetzt werden [102].

Benchmarks, die LLMs im medizinischen Bereich evaluieren, sind die in den Studien vorgestellten Benchmarks BLUE [66], BLURB [37] und MultiMedQA [75]. Da keines der Benchmarks die Aufgabe der Erstellung Vor-Anamnese bewertet, muss überprüft werden, ob die Benchmarks für diese Evaluation geeignet sind.

Das Benchmark MultiMedQA verwendet wie bereits erwähnt sieben Datensätze, die auf die Beantwortung von medizinischen Fragen zugeschnitten sind. Die medizinischen Fragen decken verschiedene medizinische Bereiche ab. Es handelt sich um Datensätze mit Fragen aus medizinischen Prüfungen, wie dem USMLE, aber auch Fragen aus medizinischer Forschung und medizinischen Verbraucherfragen. Daher eignet sich dieses Benchmark besonders gut, um die Fähigkeit Fragen im medizinischen Bereich zu beantworten zu evaluieren [75]. Da aber ein LLM für Zusammenfassung eines Vor-Anamnesegespräches gesucht wird, ist dieses Benchmark nicht geeignet.

BLUE verwendet für die Evaluation unter anderem Datensätze aus der klinischen Anwendung, also auch Datensätze, die auf PubMed Artikeln basieren und die für biomedizinische Anwendungen eingesetzt werden. Die Datensätze aus der klinischen Anwendung sind MedNLI, MedSTS, i2b2 2010 und ShARE/CLEF. Bei MedNLI handelt es sich um eine Sammlung von Satzpaaren aus MIMIC-III, die für die Validierung der Schlussfolgerungsaufgabe eingesetzt werden. MedSTS ist ebenfalls eine Sammlung von Satzpaaren, diese wurden allerdings von Experten auf die Ähnlichkeit der Textsemantik bewertet. MedSTS wird eingesetzt, um zu überprüfen wie gut die Fähigkeit ist, eingebettete Informationen zu extrahieren und redundante Informationen zu erkennen. Der Datensatz i2b2 2010 wird für die Ausgabe der Extraktion von Beziehungen eingesetzt und ist eine Sammlung von Dokumenten. Diese enthalten acht verschiedene Beziehungen zwischen Beschwerden oder Symptomen und den Behandlungen. Der letzte Datensatz ist ShARe/CLEF und wird im Aufgabenbereich NER verwendet. Der Datensatz beinhaltet eine Sammlung von nicht identifizierten freien Textnotizen von MIMIC-II. Wie bereits erwähnt, stammen die klinischen und biomedizinschen Datensätze aus den folgenden Aufgabenbereichen NER, Extraktion von Beziehungen, Satzähnlichkeit, Dokumentenklassifikation und der Schluss-

folgerungsaufgabe [66]. Im Gegensatz zu MultiMedQA ist BLUE schon eher geeignet da aus der klinischen und biomedizinischen Anwendung Datensätze verwendet werden. Aber auch bei diesem Benchmark fehlt die Bewertung des Vor-Anamnesegespräches.

Im Gegensatz zu BLUE werden in dem Benchmark BLURB nur Datensätze, die auf PubMed Artikeln basieren und für die biomedizinische Anwendung gedacht sind, verwendet. BLURB betrachtet keine Datensätze aus der klinischen Domäne, allerdings unterscheiden sich die klinischen Notizen wesentlich von den Artikeln aus der biomedizinischen Domäne. Die Aufgabenbereiche, die in BLURB betrachtet werden, sind die folgenden NER, PICO, Satzähnlichkeit, Dokumentenklassifikation und Antworten auf Fragen [37]. Da BLURB nur die Datensätze der biomedizinischen Anwendungen betrachtet, fehlen neben der Bewertung des Vor-Anamnesegespräches, die Datensätze aus der klinischen Anwendung und BLURB ist somit nicht geeignet für die Evaluation.

5.3 Evaluation

Die im medizinischen Bereich existierenden Benchmarks sind alle nicht für die Evaluation eines LLM, dass eine Vor-Anamnese erstellt, geeignet. Da in allen Benchmarks unter anderem das wichtigste Kriterium, die Qualität des Vor-Anamnesegesprächs, nicht bewertet wird und somit nicht bestimmt werden kann, wie gut das Anamnesegespräch war. Abgesehen davon, dass die Benchmarks für eine Evaluation erweitert werden müssten, sind auch die Datensätze, mit denen evaluiert wird, ein Problem. Denn das Vor-Anamnesegespräch wird auf der einen Seite nicht wie klinische Notizen und auf der anderen Seite auch nicht wie ein wissenschafticher PubMed Artikel geschrieben sein. Das Vor-Anamnesegespräch wird eher einer menschlichen Kommunikation gleichen, benötigt aber trotzdem im Hintergrund das medizinische Verständnis. Da die Benchmarks für die Textzusammenfassung, nicht das medizinische Verständnis bewerten, sind diese nicht ausreichend. Somit kann mit dem jetzigen Forschungsstand keine fundierte Auswahl getroffen werden.

Eine weitere Schwierigkeit, welches die Benchmarks MultiMedQA und BLUE betrifft, ist die Zugänglichkeit. Denn bei den Benchmarks wurden nicht die Kriterien der freien Software vorrausgesetzt. Das Benchmark MultiMedQA ist nicht für die Öffentlichkeit, Forscher oder private Personen zugänglich, somit kann es sowohl nicht für eine Evaluation innerhalb dieser Arbeit verwendet werden, als auch nicht im Allgemeinen. Bei BLUE ist die Zugänglichkeit nur bei einigen klinischen Datensätzen schwierig. Bei den Datensätzen MedNLI und ShARe/CLEF wird zusätzliches Training mit dem Namen „Data or

Specimens Only Research“ benötigt [5], um Zugriff zu erhalten. Der Datensatz MedSTS erlaubt momentan keinen Zugriff auf die „National NLP Clinical Challenges (n2c2)“ [2]. Bei i2b2 2010 ist nicht der Zugang problematisch, sondern der Datensatz. Denn von dem anfangs verwendeten Datensatz wurden Teile entfernt die nicht mehr zur Verfügung stehen. Modelle, die mit dem ganzen Datensatz evaluiert wurden, sollten nur mit Bedacht mit Modellen, die mit einer Hälfte des Datensatzes evaluiert wurden, verglichen werden, da nicht spezifiziert wurde, welche Daten aus dem Datensatz entnommen wurden [1].

6 Schluss

In dieser explorativen Studie wurde festgestellt, dass für LLMs, die eine Zusammenfassung eines Vor-Anamnesegespräch erstellen sollen, noch keine Kriterien für die Bewertung existieren. Daher wurden Kriterien entwickelt. Diese beinhalten die Qualität des Vor-Anamnesegesprächs, das medizinische Verständnis, die Qualität der Zusammenfassung und das Erinnerungsvermögen. Um die Kriterien zu überprüfen, gibt es die folgenden Benchmarks BLEU, ROUGE, BERT, BLUE, BLURB und MultiMedQA. Aber keines der vorgestellten Benchmarks kann verwendet werden, da auf der einen Seite BLEU, ROUGE und BERT nur die Fähigkeit der Zusammenfassung, ohne medizinischen Hintergrund beurteilen und BLUE, BLURB und MultiMedQA auf der anderen Seite nur das medizinische Verständnis beurteilen. Aber kein Benchmark bewertet das wichtigste Kriterium: die Qualität der Vor-Anamnese.

Um die Qualität der Vor-Anamnese bewerten zu können, müssen die Benchmarks erweitert werden. Für diese Erweiterung benötigt es Datensätze. Auf Grund einer Forschungslücke, die die KI im Anamnesebereich betrifft und der durch die Datenschutzgesetze erschwerten Beschaffung von Gesundheitsdaten, existieren solche Datensätze zur Zeit nicht. Die Datensätze für die Evaluation müssen so aufgebaut sein, das ein Vor-Anamnesegespräch oder eine Mitschrift aus einem Anamnesegespräch, von einem LLM zusammengefasst wird. Diese Zusammenfassung muss dann mit einem ausgefüllten Fragebogen oder einer von Experten zusammengefassten Version verglichen werden um festzustellen, ob alle wichtigen Informationen enthalten sind. Aber auch für das Training von LLMs in dem Bereich der Vor-Anamnese fehlen Datensätze, diese könnten ausgefüllte und unausgefüllte Anamnesebögen enthalten. Wie auch schon in der Studie „Large Language Models encode clinical knowledge“ [75] angemerkt wurde, ist in dieser Untersuchung nochmal die Notwendigkeit für Benchmarks, die reale-klinische Arbeitsabläufe abbilden, bestätigt worden. Ein weiterer Punkt, der ebenfalls in der Studie bemängelt wurde, ist die Notwendigkeit von mehrsprachigem Training und Datensätzen. Denn auch wenn die LLMs für die Zusammenfassung des Vor-Anamnesegesprächs hätten evaluiert

werden können, wäre der Einsatzbereich nur im englischsprachigem Raum möglich, da das Training und die Benchmarks mit englischen Datensätzen arbeiten.

Ein Punkt, der zwar nicht evaluiert werden kann, aber für den Einsatz von Software im medizinischen Bereich wichtig ist, ist die Frage, ob das Produkt, in diesem Fall das LLM, ein Medizinprodukt ist. Ein LLM, das mit Hilfe eines Vor-Anamnesegespräch Symptome und Beschwerden zusammenfasst, ist ein Medizinprodukt. Zwar wird am Ende noch mit einem Arzt über die Zusammenfassung der Vor-Anamnese gesprochen, trotzdem handelt es sich nicht um die reine Datensammlung, sondern um aufbereitete Daten mit indirektem Einfluss auf die Diagnose. Somit handelt es sich um ein Medizinprodukt der Klasse I mit einem niedrigen Risiko.

Auch wenn noch einige Bereiche in der Forschung behandelt werden müssen, besteht nach Betrachtung der LLMs, die bisher im medizinischen Bereich eingesetzt werden, das Potenzial ein LLM für das Erstellen einer Zusammenfassung von Symptomen, Beschwerden, eventuellen Vorerkrankungen und Medikamenten aus einem Vor-Anamnesegesprächs einzusetzen. Dafür ist neben der Forschung im Anamnesebereich wichtig, dass allgemein etablierte Standards entwickelt werden, die sowohl die Bewertung als auch den Einsatz von LLMs einschließt. Ein Schritt in die Richtung der einfacheren Bewertung ist das neue Gesetz das festlegt, dass KI ihre Trainingsdaten detaillierter offen legen soll, dies ermöglich eine bessere Beurteilung ob ein LLM für eine Aufgabe geeignet ist [19]. Abschließend dürfen vor allem für den medizinischen Bereich die Herausforderung der Halluzinationen und des Bias nicht vergessen werden. Auch die Beschaffung von Daten im Gesundheitsbereich ist schwierig und erfordert Standards.

Literaturverzeichnis

[1] *Data Sets*. Online. – URL https://n2c2.dbmi.hms.harvard.edu/data-sets. – Zuletzt besucht am 20.12.2023

[2] *n2c2 2019 — Track 1: n2c2/OHNLP Track on Clinical Semantic Textual Similarity*. Online. – URL https://portal.dbmi.hms.harvard.edu/projects/n2c2-2019-t1/. – zuletzt besucht am 20.12.2023

[3] *BigScience RAIL License v1.0*. Online. Mai 2022. – URL https://huggingface.co/spaces/bigscience/license. – zuletzt besucht am 14.12.2023

[4] 2023, Ada Health G.: *Pass auf Dich auf – mit Ada*. Online. – URL https://ada.com/de/app/. – zuletzt besucht am 09.11.2023

[5] Alistair Johnson, Roger M.: *MIMIC-III Clinical Database*. Online. September 2016. – URL https://physionet.org/content/mimiciii/1.4/. – zuletzt besucht am 20.12.2023

[6] Almazrouei, Ebtesam ; Alobeidli, Hamza ; Alshamsi, Abdulaziz ; Cappelli, Alessandro ; Cojocaru, Ruxandra ; Debbah, Mérouane ; Goffinet Étienne ; Hesslow, Daniel ; Launay, Julien ; Malartic, Quentin ; Mazzotta, Daniele ; Noune, Badreddine ; Pannier, Baptiste ; Penedo, Guilherme: *The Falcon Series of Open Language Models*. 2023

[7] Alsentzer, Emily ; Murphy, John ; Boag, William ; Weng, Wei-Hung ; Jindi, Di ; Naumann, Tristan ; McDermott, Matthew: Publicly Available Clinical BERT Embeddings. In: Rumshisky, Anna (Hrsg.) ; Roberts, Kirk (Hrsg.) ; Bethard, Steven (Hrsg.) ; Naumann, Tristan (Hrsg.): *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA : Association for Computational Linguistics, Juni 2019, S. 72–78. – URL https://aclanthology.org/W19-1909

[8] Anthropic: *Supported countries and regions: Claude.ai*. Online. – URL https://www.anthropic.com/claude-ai-locations. – zuletzt besucht am 10.12.2023

[9] baeldung: *Natural Language Processing: Bleu Score*. Online. März 2023. – URL https://www.baeldung.com/cs/nlp-bleu-score. – Zuletzt besucht am 18.12.2023

[10] Basyal, Lochan ; Sanghvi, Mihir: *Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models*. Oktober 2023

[11] Beltagy, Iz ; Lo, Kyle ; Cohan, Arman: SciBERT: A Pretrained Language Model for Scientific Text. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, November 2019, S. 3615–3620

[12] Benedetti, Alessandro: *How to Choose the Right Large Language Model for Your Domain – Open Source Edition*. Online. Juni 2023. – URL https://sease.io/2023/06/how-to-choose-the-right-large-language-model-for-your-domain-open-source-edition.html. – zuletzt besucht am 15.09.2023

[13] Betriebssystem, GNU: *Freie Software. Was ist das?* Online. 2019. – URL https://www.gnu.org/philosophy/free-sw.de.html#:~:text=FreieSoftwareistSoftware,die,zuÃd'ndernundzuverbessern.. – zuletzt besucht am 09.12.2023

[14] BigScience: *Bloom-Z*. Online. – URL https://huggingface.co/bigscience/bloomz. – zuletzt besucht am 14.12.2023

[15] Black, Sid ; Biderman, Stella ; Hallahan, Eric ; Anthony, Quentin ; Gao, Leo ; Golding, Laurence ; He, Horace ; Leahy, Connor ; McDonell, Kyle ; Phang, Jason ; Pieler, Michael ; Prashanth, USVSN S. ; Purohit, Shivanshu ; Reynolds, Laria ; Tow, Jonathan ; Wang, Ben ; Weinbach, Samuel: *GPT-NeoX-20B: An Open-Source Autoregressive Language Model*. 2022

[16] Bordoloi, Pritam: *Why Meta Took Down its 'Hallucinating' AI Model Galactica?* Online. November 2022. – URL https://analyticsindiamag.com/why-

meta-took-down-its-hallucinating-ai-model-galactica/. – Zuletzt besucht am 12.12.2023

[17] Brautzsch, Jessica: *Wie KI-Forschung in der Medizin trotz Datenschutz gelingen kann.* Online. Juni 2021. – URL https://www.mdr.de/wissen/kuenstliche-intelligenz-medizin-datenschutz-100.html. – Zuletzt besucht am 16.12.2023

[18] CalderaAI: *CalderaAI/30B-Lazarus.* Online. – URL https://huggingface.co/CalderaAI/30B-Lazarus. – Zuletzt besucht am 03.12.2023

[19] Carsten Volkery, Larissa Holzki und Josefine F.: *EU beschließt umfangreichstes KI-Gesetz der Welt – das sind die wichtigsten Punkte.* Online. Dezember 2023. – URL https://www.handelsblatt.com/politik/international/ai-act-eu-beschliesst-umfangreichstes-ki-gesetz-der-welt-das-sind-die-wichtigsten-punkte/100002256.html. – Zuletzt besucht am 20.12.2023

[20] Chandrakant, Kumar: *Introduction to Large Language Models.* Online. Juli 2023. – URL https://www.baeldung.com/cs/large-language-models. – zuletzt besucht am 13.09.2023

[21] chrissikraus: *Was ist Closed Source?* Online. März 2019. – URL https://www.it-business.de/was-ist-closed-source-a-92d49435c24ca7762592751c6068faac/. – zuletzt besucht am 20.12.2023

[22] Christoph Auer, Matthias R.: *Gesundheit digital: Perspektiven zur Digitalisierung im Gesundheitswesen.* Kap. Künstliche Intelligenz im Gesundheitswesen, S. 33–46, Springer Berlin Heidelberg, 2019. – ISBN 9783662576113

[23] Chung, Hyung W. ; Hou, Le ; Longpre, Shayne ; Zoph, Barret ; Tay, Yi ; Fedus, William ; Li, Yunxuan ; Wang, Xuezhi ; Dehghani, Mostafa ; Brahma, Siddhartha ; Webson, Albert ; Gu, Shixiang S. ; Dai, Zhuyun ; Suzgun, Mirac ; Chen, Xinyun ; Chowdhery, Aakanksha ; Castro-Ros, Alex ; Pellat, Marie ; Robinson, Kevin ; Valter, Dasha ; Narang, Sharan ; Mishra, Gaurav ; Yu, Adams ; Zhao, Vincent ; Huang, Yanping ; Dai, Andrew ; Yu, Hongkun ; Petrov, Slav ; Chi, Ed H. ; Dean, Jeff ; Devlin, Jacob ; Roberts, Adam ; Zhou, Denny ; Le, Quoc V. ; Wei, Jason: *Scaling Instruction-Finetuned Language Models.* 2022

[24] Computer, Together: *OpenChatKit: An Open Toolkit and Base Model for Dialogue-style Applications*. 3 2023. – URL https://github.com/togethercomputer/OpenChatKit

[25] Conover, Mike ; Hayes, Matt ; Mathur, Ankit ; Xie, Jianwei ; Wan, Jun ; Shah, Sam ; Ghodsi, Ali ; Wendell, Patrick ; Zaharia, Matei ; Xin, Reynold: *Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM*. 2023. – URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm. – Zugriffsdatum: 2023-06-30

[26] Dettmers, Tim ; Pagnoni, Artidoro ; Holtzman, Ari ; Zettlemoyer, Luke: *QLoRA: Efficient Finetuning of Quantized LLMs*. 2023

[27] Dictonary, Cambridge: *state-of-the-art*. Online. – URL https://dictionary.cambridge.org/de/worterbuch/englisch/state-of-the-art. – Zuletzt besucht am 20.12.2023

[28] Dilmegani, Cem: *Large Language Model Training in 2023*. Online. Februar 2023. – URL https://research.aimultiple.com/large-language-model-training/. – Besucht am 03.06.2023

[29] Du, Zhengxiao ; Qian, Yujie ; Liu, Xiao ; Ding, Ming ; Qiu, Jiezhong ; Yang, Zhilin ; Tang, Jie: *GLM: General Language Model Pretraining with Autoregressive Blank Infilling*. 2021

[30] Face, Hugging: *FLAN-T5*. Online. – URL https://huggingface.co/docs/transformers/main/en/model_doc/flan-t5. – Zuletzt besucht am 01.12.2023

[31] Face, Hugging: *google/ul2*. Online. – URL https://huggingface.co/google/ul2. – Zuletzt besucht am 01.12.2023

[32] Face, Hugging: *Sentence Similarity*. Online. – URL https://huggingface.co/tasks/sentence-similarity. – zuletzt besucht am 11.12.2023

[33] Face, Hugging: *T5*. Online. – URL https://huggingface.co/docs/transformers/model_doc/t5. – zuletzt besucht am 20.11.2023

[34] Ferrara, Emilio: Should ChatGPT be biased? Challenges and risks of bias in large language models. In: *First Monday* (2023), November. – ISSN 1396-0466

[35] Gesundheitswesen (IQWiG), Institut für Qualität und Wirtschaftlichkeit im: *Blinddarmentzündung (Appendizitis)*. Online. November 2021. – URL https://www.gesundheitsinformation.de/blinddarmentzuendung-appendizitis.html. – Zuletzt besucht am 18.12.2023

[36] Ghahramani, Zoubin: *Introducing PaLM 2*. Online. Mai 2023. – URL https://blog.google/technology/ai/google-palm-2-ai-large-language-model/. – zuletzt besucht am 08.11.2023

[37] Gu, Yu ; Tinn, Robert ; Cheng, Hao ; Lucas, Michael ; Usuyama, Naoto ; Liu, Xiaodong ; Naumann, Tristan ; Gao, Jianfeng ; Poon, Hoifung: Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. In: *ACM Transactions on Computing for Healthcare* 3 (2021), Oktober, Nr. 1, S. 1–23. – ISSN 2637-8051

[38] Gunasekar, Suriya ; Zhang, Yi ; Aneja, Jyoti ; Mendes, Caio César T. ; Giorno, Allie D. ; Gopi, Sivakanth ; Javaheripi, Mojan ; Kauffmann, Piero ; Rosa, Gustavo de ; Saarikivi, Olli ; Salim, Adil ; Shah, Shital ; Behl, Harkirat S. ; Wang, Xin ; Bubeck, Sébastien ; Eldan, Ronen ; Kalai, Adam T. ; Lee, Yin T. ; Li, Yuanzhi: *Textbooks Are All You Need*. 2023

[39] Health, Glass: *Glass AI Beta*. Online. – URL https://glass.health/ai. – Besucht am 21.06.2023

[40] Hosni, Youssef: *Top 10 Open Source LLMs To USE In Your Next LLM Application*. Online. August 2023. – URL https://pub.towardsai.net/top-10-open-source-llms-to-use-in-your-next-llm-application-fbfc51542b78. – zuletzt besucht am 10.09.2023

[41] HuggingFace: *BERT*. Online. – URL https://huggingface.co/docs/transformers/model_doc/bert. – Zuletzt besucht am 01.12.2023

[42] Höfel; Michael Steinel; Dr. Frank Antwerpes; David Ekert; Bijan Fink; Dr. No, Natascha van den: *PICO-Schema*. Online. Juni 2022. – URL https://flexikon.doccheck.com/de/PICO-Schema. – Zuletzt besucht am 17.12.2023

[43] Iyer, Srinivasan ; Lin, Xi V. ; Pasunuru, Ramakanth ; Mihaylov, Todor ; Simig, Daniel ; Yu, Ping ; Shuster, Kurt ; Wang, Tianlu ; Liu, Qing ; Koura, Punit S. ; Li, Xian ; O'Horo, Brian ; Pereyra, Gabriel ; Wang, Jeff ; Dewan, Christopher ;

Celikyilmaz, Asli ; Zettlemoyer, Luke ; Stoyanov, Ves: *OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization.* 2022

[44] Kaddour, Jean ; Harris, Joshua ; Mozes, Maximilian ; Bradley, Herbie ; Raileanu, Roberta ; McHardy, Robert: Challenges and Applications of Large Language Models. In: *arXiv preprint arXiv:2307.10169* (2023)

[45] Kung, Tiffany H. ; Cheatham, Morgan ; Medenilla, Arielle ; Sillos, Czarina ; Leon, Lorie D. ; Elepaño, Camille ; Madriaga, Maria ; Aggabao, Rimel ; Diaz-Candido, Giezel ; Maningo, James ; Tseng, Victor: Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. In: *Plos Digital Health* 2 (2023), S. e0000198. – ISSN 2767-3170

[46] Lee, Jinhyuk ; Yoon, Wonjin ; Kim, Sungdong ; Kim, Donghyeon ; Kim, Sunkyu ; So, Chan H. ; Kang, Jaewoo: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. In: *Bioinformatics* 36 (2019), September, Nr. 4, S. 1234–1240. – ISSN 1367-4811

[47] Levine, Xinyang Geng; Arnav Gudibande; Hao Liu; Eric Wallace; Pieter Abbeel; S. ; Song, Dawn: *Koala: A Dialogue Model for Academic Research.* Online. April 2023. – URL https://bair.berkeley.edu/blog/2023/04/03/koala/. – Zuletzt besucht am 16.12.2023

[48] Liu, Tiedong ; Low, Bryan Kian H.: *Goat: Fine-tuned LLaMA Outperforms GPT-4 on Arithmetic Tasks.* 2023

[49] Luber, Dipl.-Ing. (FH) S.: *Was ist ein Large Language Model (LLM)?* Online. Mai 2023. – URL https://www.cloudcomputing-insider.de/was-ist-ein-large-language-model-llm-a-9b7bdd0c3766b5a9c0ee1e0c909790a3/#:~:text=DasAkronymlautetLLM.,verwendenDeep-Learning-Algorithmen.. – zuletzt besucht am 20.09.2023

[50] Luo, Renqian ; Sun, Liai ; Xia, Yingce ; Qin, Tao ; Zhang, Sheng ; Poon, Hoifung ; Liu, Tie-Yan: BioGPT: generative pre-trained transformer for biomedical text generation and mining. In: *Briefings in Bioinformatics* 23 (2022), September, Nr. 6. – URL http://dx.doi.org/10.1093/bib/bbac409. – ISSN 1477-4054

[51] Luo, Ziyang ; Xu, Can ; Zhao, Pu ; Sun, Qingfeng ; Geng, Xiubo ; Hu, Wenxiang ; Tao, Chongyang ; Ma, Jing ; Lin, Qingwei ; Jiang, Daxin: WizardCoder: Empowering Code Large Language Models with Evol-Instruct. In: *arXiv preprint arXiv:2306.08568* (2023)

[52] Lutkevich, Ben: *16 of the best large language models.* Online. Oktober 2023. – URL https://www.techtarget.com/whatis/feature/12-of-the-best-large-language-models. – zuletzt besucht am 10.11.2023

[53] MacFarland, Alex: *Best Open Source LLMs.* Online. August 2023. – URL https://www.unite.ai/best-open-source-llms/. – zuletzt besucht am 10.09.2023

[54] Marshall, Christopher: *What is named entity recognition (NER) and how can I use it?* Online. Dezember 2019. – URL https://medium.com/mysuperai/what-is-named-entity-recognition-ner-and-how-can-i-use-it-2b68cf6f545d. – zuletzt besucht am 05.12.2023

[55] Medizinprodukte, Bundesinstitut für Arzneimittel und: *Abgrenzung und Klassifizierung.* Online. – URL https://www.bfarm.de/DE/Medizinprodukte/Aufgaben/Abgrenzung-und-Klassifizierung/_node.html. – Zuletzt besucht am 20.12.2023

[56] Mesri, Alparslan: *Text Summarization: How To Calculate Rouge Score.* Online. August 2023. – URL https://medium.com/mlearning-ai/text-summarization-84ada711c49c. – Zuletzt besucht am 18.12.2023

[57] Meta: *LLaMA2.* Online. – URL https://ai.meta.com/llama/. – Zuletzt besucht am 01.12.2023

[58] Mitra, Arindam ; Corro, Luciano D. ; Mahajan, Shweti ; Codas, Andres ; Simoes, Clarisse ; Agrawal, Sahaj ; Chen, Xuxi ; Razdaibiedina, Anastasia ; Jones, Erik ; Aggarwal, Kriti ; Palangi, Hamid ; Zheng, Guoqing ; Rosset, Corby ; Khanpour, Hamed ; Awadallah, Ahmed: *Orca 2: Teaching Small Language Models How to Reason.* 2023

[59] Muennighoff, Margaret Mitchell; Giada Pistilli; Yacine Jernite; Ezinwanne Ozoani; Marissa Gerchick; Nazneen Rajani; Sasha Luccioni; Irene Solaiman; Maraim Masoud; Somaieh Nikpoor; Carlos Muñoz Ferrandis; Stas Bekman; Christopher Aki-

ki; Danish Contractor; David Lansky; Angelina McMillan-Major; Tristan Thrush; Suzana Ilić; Gérard Dupont; Shayne Longpre; Manan Dey; Stella Biderman; Douwe Kiela; Emi Baylor; Teven Le Scao; Aaron Gokaslan; Julien Launay; N.: *Bloom*. Online. – URL https://huggingface.co/bigscience/bloom. – Zuletzt besucht am 10.12.2023

[60] Naveed, Humza ; Khan, Asad U. ; Qiu, Shi ; Saqib, Muhammad ; Anwar, Saeed ; Usman, Muhammad ; Akhtar, Naveed ; Barnes, Nick ; Mian, Ajmal: *A Comprehensive Overview of Large Language Models*. 2023

[61] Neupane, Parlad: *Understanding Text Classification in NLP with Movie Review Example*. Online. August 2023. – URL https://www.analyticsvidhya.com/blog/2020/12/understanding-text-classification-in-nlp-with-movie-review-example-example/#h-what-is-text-classification. – Zuletzt besucht am 17.12.2023

[62] Nijkamp, Erik ; Pang, Bo ; Hayashi, Hiroaki ; Tu, Lifu ; Wang, Huan ; Zhou, Yingbo ; Savarese, Silvio ; Xiong, Caiming: *CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis*. 2023

[63] Omiye, Jesutofunmi A. ; Gui, Haiwen ; Rezaei, Shawheen J. ; Zou, James ; Daneshjou, Roxana: *Large language models in medicine: the potentials and pitfalls*. 2023

[64] Patel, Ruchi ; Tanwani, Sanjay ; Patidar, Chhaya: Relation Extraction between Medical Entities using Deep Learning Approach. In: *Informatica* 45 (2021), September, Nr. 3. – ISSN 0350-5596

[65] Patz, Laura: Manche Aufgaben kann KI besser als wir. In: *Apothekenumschau* 08B (2023), August, S. 18–20

[66] Peng, Yifan ; Yan, Shankai ; Lu, Zhiyong: Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In: Dmner-Fushman, Dina (Hrsg.) ; Cohen, Kevin B. (Hrsg.) ; Ananiadou, Sophia (Hrsg.) ; Tsujii, Junichi (Hrsg.): *Proceedings of the 18th BioNLP Workshop and Shared Task*, Association for Computational Linguistics, August 2019, S. 58–65

[67] Public Health Infrastructure Center, Public Health Law P.: *Health Insurance Portability and Accountability Act of 1996 (HIPAA)*. Online. Juni

2022. – URL https://www.cdc.gov/phlp/publications/topic/hipaa.html. – Zuletzt besucht am 18.12.2023

[68] Rowley, Jason D.: *LLM Benchmarks: Guide to Evaluating Language Models*. Online. Juni 2023. – URL https://deepgram.com/learn/llm-benchmarks-guide-to-evaluating-language-models#. – Zuletzt besucht am 20.12.2023

[69] Sahota, Harpreet: *Top 10 List of Large Language Models Reshaping the Open-Source Arena in 2023*. Online. August 2023. – URL https://deci.ai/blog/list-of-large-language-models-in-open-source/. – zuletzt besucht am 10.09.2023

[70] Sanh, Victor ; Webson, Albert ; Raffel, Colin ; Bach, Stephen H. ; Sutawika, Lintang ; Alyafeai, Zaid ; Chaffin, Antoine ; Stiegler, Arnaud ; Scao, Teven L. ; Raja, Arun ; Dey, Manan ; Bari, M S. ; Xu, Canwen ; Thakker, Urmish ; Sharma, Shanya S. ; Szczechla, Eliza ; Kim, Taewoon ; Chhablani, Gunjan ; Nayak, Nihal ; Datta, Debajyoti ; Chang, Jonathan ; Jiang, Mike Tian-Jian ; Wang, Han ; Manica, Matteo ; Shen, Sheng ; Yong, Zheng X. ; Pandey, Harshit ; Bawden, Rachel ; Wang, Thomas ; Neeraj, Trishala ; Rozen, Jos ; Sharma, Abheesht ; Santilli, Andrea ; Fevry, Thibault ; Fries, Jason A. ; Teehan, Ryan ; Bers, Tali ; Biderman, Stella ; Gao, Leo ; Wolf, Thomas ; Rush, Alexander M.: *Multitask Prompted Training Enables Zero-Shot Task Generalization*. 2022

[71] Science, Data: *Document Classification With Machine Learning: Computer Vision, OCR, NLP, and Other Techniques*. Online. November 2021. – URL https://www.altexsoft.com/blog/document-classification/. – Zuletzt besucht am 19.12.2023

[72] Sha, Arjun: *12 Best Large Language Models (LLMs) in 2023*. Online. Juni 2023. – URL https://beebom.com/best-large-language-models-llms/. – zuletzt besucht am 10.11.2023

[73] Shane, Janelle: *Künstliche Intelligenz - Wie sie funktioniert und wann sie scheitert : Eine unterhaltsame Reise in die seltsame Welt der Algorithmen, neuronalen Netze und versteckten Giraffen*. Kap. Kapitel 1: Was ist KI?, S. 17–35, O'Reilly, 2021. – ISBN 978-3-96009-160-8

[74] Sharma, Prashant: *Dependency Parsing in Natural Language Processing with Examples*. Online. August 2023. – URL https://www.analyticsvidhya.com/blog/2021/12/dependency-parsing-in-natural-language-processing-with-examples/. – Zuletzt besucht am 18.12.2023

[75] Singhal, Karan ; Azizi, Shekoofeh ; Tu, Tao ; Mahdavi, S. S. ; Wei, Jason ; Chung, Hyung W. ; Scales, Nathan ; Tanwani, Ajay ; Cole-Lewis, Heather ; Pfohl, Stephen ; Payne, Perry ; Seneviratne, Martin ; Gamble, Paul ; Kelly, Chris ; Babiker, Abubakr ; Schärli, Nathanael ; Chowdhery, Aakanksha ; Mansfield, Philip ; Demner-Fushman, Dina ; Arcas, Blaise A. y ; Webster, Dale ; Corrado, Greg S. ; Matias, Yossi ; Chou, Katherine ; Gottweis, Juraj ; Tomasev, Nenad ; Liu, Yun ; Rajkomar, Alvin ; Barral, Joelle ; Semturs, Christopher ; Karthikesalingam, Alan ; Natarajan, Vivek: Large language models encode clinical knowledge. In: *Nature* 620 (2023), jul, Nr. 7972, S. 172–180

[76] Spiegel: *Arbeitnehmer loben telefonische Krankschreibung*. Online. Dezember 2021. – URL https://www.spiegel.de/wirtschaft/arbeitnehmer-loben-telefonische-krankschreibung-a-bde77c3f-7b2c-47b6-a53d-ae81d4ea1455. – zuletzt besucht: 23.10.2023

[77] StabilityAI: *stabilityai/StableBeluga-7B*. Online. – URL https://huggingface.co/stabilityai/StableBeluga-7B. – Zuletzt besucht am 15.12.2023

[78] StabilityAI: *Stable LM*. Online. – URL https://github.com/Stability-AI/StableLM. – Zuletzt besucht am 17.12.2023

[79] Stallman, Richard: *Warum „Open Source“ das Ziel Freie Software verfehlt*. Online. Januar 2020. – URL https://www.gnu.org/philosophy/open-source-misses-the-point. – zuletzt besucht am 09.12.2023

[80] Taori, Rohan ; Gulrajani, Ishaan ; Zhang, Tianyi ; Dubois, Yann ; Li, Xuechen ; Guestrin, Carlos ; Liang, Percy ; Hashimoto, Tatsunori B.: *Stanford Alpaca: An Instruction-following LLaMA model*. https://github.com/tatsu-lab/stanford_alpaca. 2023

[81] Taylor, Ross ; Kardas, Marcin ; Cucurull, Guillem ; Scialom, Thomas ; Hartshorn, Anthony ; Saravia, Elvis ; Poulton, Andrew ; Kerkez, Viktor ; Stojnic, Robert: *Galactica: A Large Language Model for Science*. 2022

[82] Team, The MosaicML N.: *Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs*. Online. Mai 2023. – URL https://www.mosaicml.com/blog/mpt-7b. – Zuletzt besucht am 02.12.2023

[83] Thirunavukarasu, Arun J. ; Ting, Darren Shu J. ; Elangovan, Kabilan ; Gutierrez, Laura ; Tan, Ting F. ; Ting, Daniel Shu W.: Large language models in medicine. In: *Nature Medicine* 29 (2023), Juli, Nr. 8, S. 1930–1940. – ISSN 1546-170X

[84] Touvron, Hugo ; Lavril, Thibaut ; Izacard, Gautier ; Martinet, Xavier ; Lachaux, Marie-Anne ; Lacroix, Timothée ; Rozière, Baptiste ; Goyal, Naman ; Hambro, Eric ; Azhar, Faisal ; Rodriguez, Aurelien ; Joulin, Armand ; Grave, Edouard ; Lample, Guillaume: *LLaMA: Open and Efficient Foundation Language Models*. 2023

[85] Tustumi, Francisco ; Andreollo, Nelson A. ; Aguilar-Nascimento, José E. de: *FUTURE OF THE LANGUAGE MODELS IN HEALTHCARE: THE ROLE OF CHATGPT*. 2023

[86] Vaswani, Ashish ; Shazeer, Noam ; Parmar, Niki ; Uszkoreit, Jakob ; Jones, Llion ; Gomez, Aidan N. ; Kaiser, Lukasz ; Polosukhin, Illia: Attention Is All You Need. In: *arXiv preprint arXiv:1706.03762* (2017). – Version 7: 2. Aug. 2023

[87] Wang, Haochun ; Liu, Chi ; Xi, Nuwa ; Qiang, Zewen ; Zhao, Sendong ; Qin, Bing ; Liu, Ting: *HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledge*. 2023

[88] Wang, Haochun ; Liu, Chi ; Xi, Nuwa ; Qiang, Zewen ; Zhao, Sendong ; Qin, Bing ; Liu, Ting: *HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledge*. 2023

[89] Wang, Yizhong ; Mishra, Swaroop ; Alipoormolabashi, Pegah ; Kordi, Yeganeh ; Mirzaei, Amirreza ; Arunkumar, Anjana ; Ashok, Arjun ; Dhanasekaran, Arut S. ; Naik, Atharva ; Stap, David ; Pathak, Eshaan ; Karamanolakis, Giannis ; Lai, Haizhi G. ; Purohit, Ishan ; Mondal, Ishani ; Anderson, Jacob ; Kuznia, Kirby ; Doshi, Krima ; Patel, Maitreya ; Pal, Kuntal K. ; Moradshahi, Mehrad ; Parmar, Mihir ; Purohit, Mirali ; Varshney, Neeraj ; Kaza, Phani R. ; Verma, Pulkit ; Puri, Ravsehaj S. ; Karia, Rushang ; Sampat, Shailaja K. ; Doshi, Savan ; Mishra, Siddhartha ; Reddy, Sujan ; Patro, Sumanta ; Dixit, Tanay ; Shen, Xudong ; Baral, Chitta ; Choi, Yejin ; Smith,

Noah A. ; Hajishirzi, Hannaneh ; Khashabi, Daniel: *Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks*. 2022

[90] Wang, Yue ; Le, Hung ; Gotmare, Akhilesh D. ; Bui, Nghi D. Q. ; Li, Junnan ; Hoi, Steven C. H.: *CodeT5+: Open Code Large Language Models for Code Understanding and Generation*. 2023

[91] Werra; Loubna Ben Allal, Leandro von: *StarCoder: A State-of-the-Art LLM for Code*. Online. Mai 2023. – URL https://huggingface.co/blog/starcoder. – Zuletzt besucht am 01.12.2023

[92] Wuttke, Laurenz: *Deep Learning: Definition, Beispiele Frameworks*. Online. – URL https://datasolut.com/was-ist-deep-learning/#Neuronale-Netze-Deep-Learning. – Zuletzt besucht am 21.12.2023

[93] Wuttke, Laurenz: *Text Mining: Definition, Methoden und Anwendung*. Online. Juni 2022. – URL https://datasolut.com/wiki/text-mining/. – Zuletzt besucht am 21.12.2023

[94] Xu, Can ; Sun, Qingfeng ; Zheng, Kai ; Geng, Xiubo ; Zhao, Pu ; Feng, Jiazhan ; Tao, Chongyang ; Jiang, Daxin: *WizardLM: Empowering Large Language Models to Follow Complex Instructions*. 2023

[95] Yang, Xi ; Chen, Aokun ; PourNejatian, Nima ; Shin, Hoo C. ; Smith, Kaleb E. ; Parisien, Christopher ; Compas, Colin ; Martin, Cheryl ; Flores, Mona G. ; Zhang, Ying ; Magoc, Tanja ; Harle, Christopher A. ; Lipori, Gloria ; Mitchell, Duane A. ; Hogan, William R. ; Shenkman, Elizabeth A. ; Bian, Jiang ; Wu, Yonghui: *GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records*. 2022

[96] Yasunaga, Michihiro ; Leskovec, Jure ; Liang, Percy: *LinkBERT: Pretraining Language Models with Document Links*. 2022

[97] Zeng, Wei ; Ren, Xiaozhe ; Su, Teng ; Wang, Hui ; Liao, Yi ; Wang, Zhiwei ; Jiang, Xin ; Yang, ZhenZhang ; Wang, Kaisheng ; Zhang, Xiaoda ; Li, Chen ; Gong, Ziyan ; Yao, Yifan ; Huang, Xinjing ; Wang, Jun ; Yu, Jianfeng ; Guo, Qi ; Yu, Yue ; Zhang, Yan ; Wang, Jin ; Tao, Hengtao ; Yan, Dasen ; Yi, Zexuan ; Peng, Fang ; Jiang, Fangqing ; Zhang, Han ; Deng, Lingfeng ; Zhang, Yehong ; Lin, Zhe ; Zhang, Chao ; Zhang, Shaojie ; Guo, Mingyue ; Gu, Shanzhi ; Fan, Gaojun ; Wang, Yaowei ; Jin, Xuefeng ; Liu, Qun ; Tian, Yonghong: *PanGu-:*

Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation. 2021

[98] Zhang, Susan ; Roller, Stephen ; Goyal, Naman ; Artetxe, Mikel ; Chen, Moya ; Chen, Shuohui ; Dewan, Christopher ; Diab, Mona ; Li, Xian ; Lin, Xi V. ; Mihaylov, Todor ; Ott, Myle ; Shleifer, Sam ; Shuster, Kurt ; Simig, Daniel ; Koura, Punit S. ; Sridhar, Anjali ; Wang, Tianlu ; Zettlemoyer, Luke: *OPT: Open Pre-trained Transformer Language Models*. 2022

[99] Zhang, Xuanyu ; Yang, Qing ; Xu, Dongliang: *XuanYuan 2.0: A Large Chinese Financial Chat Model with Hundreds of Billions Parameters*. 2023

[100] Zhang, Zhengyan ; Gu, Yuxian ; Han, Xu ; Chen, Shengqi ; Xiao, Chaojun ; Sun, Zhenbo ; Yao, Yuan ; Qi, Fanchao ; Guan, Jian ; Ke, Pei ; Cai, Yanzheng ; Zeng, Guoyang ; Tan, Zhixing ; Liu, Zhiyuan ; Huang, Minlie ; Han, Wentao ; Liu, Yang ; Zhu, Xiaoyan ; Sun, Maosong: *CPM-2: Large-scale Cost-effective Pre-trained Language Models*. 2021

[101] Zheng, Lianmin ; Chiang, Wei-Lin ; Sheng, Ying ; Zhuang, Siyuan ; Wu, Zhanghao ; Zhuang, Yonghao ; Lin, Zi ; Li, Zhuohan ; Li, Dacheng ; Xing, Eric. P. ; Zhang, Hao ; Gonzalez, Joseph E. ; Stoica, Ion: *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. 2023

[102] Özbolat, Hatice: *Text Summarization: How to Calculate BertScore*. Online. September 2023. – URL https://haticeozbolat17.medium.com/text-summarization-how-to-calculate-bertscore-771a51022964. – Zuletzt besucht am 18.12.2023

A Anhang

A.1 Überblick über existente LLMs

LLM	Freie Software
T5 [60]	ja
mT5 [60]	ja
GPT-3 [60]	nein
T0 [60]	ja
CPM-2 [60]	ja
PanGu- α [60]	ja
Codex [60]	nein
ERNIE 3.0 [60]	nein
Jurassic-1 [60]	nein
HyperCLOVA [60]	nein
Yuan 1.0 [60]	nein
Gopher [60]	nein
ERNIE 3.0 Titan [60]	nein
GLaM [60]	nein
LaMDA [60]	nein
WebGPT [60]	nein
OPT-IML [60]	ja
MT0 [60]	ja
Galactica [60]	ja
GLM [60]	ja
OPT [60]	ja
UL2 [60]	ja
Tk-Instruct [60]	ja
GPT-NeoX-20B [60]	ja

LLM	Freie Software
CodeGen [60]	ja
MT-NLG [60]	nein
AlphaCode [60]	nein
Chinchilla [60]	nein
PaLM [60]	nein
AlexaTM [60]	nein
Sparrow [60]	nein
U-PaLM [60]	nein
Flan-U-PaLM [60]	nein
BLOOM [60]	ja unter der BigScience Lizenz
ChatGPT [60]	nein
Code LlaMa [60]	ja
LLaMA 2 [60]	ja
WizardCoder [60]	ja
MPT [60]	ja
Goat [60]	ja
CodeT5+ [60]	ja
StarCoder [60]	ja
XuanYuan 2.0 [60]	ja
Koala [60]	ja
WizardLM [60]	ja
HuaTuo [60]	ja
Vicuna [60]	ja
Alpaca [60]	ja
LLaMA [60]	ja
PanGu-Σ [60]	nein
BloombergGPT [60]	nein
GPT-4 [60]	nein
Claude [60]	nein man muss dafür bezahlen
Bard [60]	nein
GPT-4 [72]	nein
GPT-3.5 [72]	nein
PaLM 2 (Bison-001) [72]	nein
Claude v1 [72]	nein
Cohere [72]	nein

LLM	Freie Software
Falcon [72]	ja
LLaMA [72]	ja
Guanaco-65B [72]	ja
Vicuna 33B [72]	ja
MPT-30B [72]	ja
30B-Lazarus [72]	ja
WizardLM [72]	ja
GPT4All [72]	ja
BERT [52]	ja
Claude [52]	nein
Cohere [52]	nein
Ernie [52]	nein
Falcon 40B [52]	ja
Galactica [52]	ja
GPT-3 [52]	nein
GPT-3.5 [52]	nein
GPT-4 [52]	nein
Lamda [52]	nein
LLaMA [52]	ja
Orca [52]	ja
PaLM [52]	nein
Phi-1 [52]	ja
StableLM [52]	ja
Vicuna 33B [52]	ja
LLAMA 2 [53]	ja
Claude 2 [53]	nein
MPT-7B [53]	ja
Falcon [53]	ja
Vicuna-13B [53]	ja
LLaMA [69]	ja
LLaMA2 [69]	ja
Alpaca [69]	ja
Vicuna [69]	ja
Guanaco [69]	ja
RedPajama [69]	nein

LLM	Freie Software
Falcon [69]	ja
FLAN-T5 [69]	ja
Stable-Beluga [69]	ja
MPT [69]	ja
LLaMA [40]	ja
Falcon [40]	ja
Dolly [40]	ja
Guanaco [40]	ja
BloomZ [40]	ja unter BigScience Lizenz
Alpaca [40]	ja
OpenChatKit [40]	ja
GPT4All [40]	ja
Vicuna [40]	ja
FLAN-T5 [40]	ja
BioGPT [63]	ja
BioMedLM [63]	ja
PubMedBERT [37]	nein
BioBERT [46]	ja
BioLinkBERT [63]	ja
ClinicalBERT [63]	ja
Flan-PaLM [23]	nein
Med-PaLM [75]	nein
PMC-LLaMA [63]	ja
GatorTron [95]	ja
SciBERT [11]	ja

Tabelle A.1: Überblick über alle LLMs aus verschiedenen Quellen

A.2 Überblick über alle freien LLM

LLM	vortrainiert	feinabgestimmt	Besonderheiten
T5 [33]	ja	nein	
mT5 [33]	ja	nein	basiert auf T5
T0 [70]	ja	nein	basiert auf T5
CPM-2 [100]	ja	nein	3-stufiges vortrainieren, 1. Chinesische Daten, 2. Bilinguale Daten, 3. multistage Daten
PanGu- α [97]	ja	nein	Chinesische Daten fürs Training
OPT-IML [43]	ja	ja	fein-abgestimmte Version von OPT
MT0	ja	ja	basiert auf mT5(als multilingual language models)
Galactica [81]	ja	nein	für scientific knowledge task, Kritik wegen Halluzinationen [16]
GLM [29]	ja	nein	pre-trained mit autoregressive blank filling objective
OPT [98]	ja	nein	evtl. kein Access
UL2 [31]	ja	ja	ähnlich zu T5, gleicher Satz tokenizer verschiedene Objektive und C4 Korpus
Tk-Instruct [89]	ja	ja	build upon T5 models
GPT-NeoX-20B [15]	ja	nein	architecture ähnelt GPT-3
CodeGen [62]	ja	nein	für Code Intelligenz
BLOOM [59]	ja	nein	Architektur ähnelt GPT-3 Architektur
Code LLaMA [57]	ja	ja	basiert auf LLaMA 2, für Code Generierung
LLaMA 2 [57]	ja	ja	für Chat-Anwendung fein-abgestimmt
WizardCoder [51]	ja		basiert auf Code LLaMA

LLM	vortrainiert	feinabgestimmt	Besonderheiten
MPT [82]	ja	nein	es existieren verschiedene fein-abgestimmte Versionen von MPT
Goat [48]	ja	ja	fein-abgestimmte LLaMA Model für synthetisch generierte Datenstrukturen (DS) oder Datensammlungen??
CodeT5+ [90]	ja		verbesserte Model Architektur für Code Intelligenz
StarCoder [91]	ja	ja	fein-abgestimmte Version von StarCoderBase trainiert auf Daten von GitHub für Code Intelligenz
XuanYuan 2.0 [99]	ja		basiert auf Bloom, Chinesischer Finanz Chat-Bot
Koala [47]	ja	ja	fein-abgestimmter Chat-Bot basierend auf LLaMA
WizardLM [94]	ja	ja	fein-abgestimmtes LLaMA Chatbot
HuaTuo [88]	ja	ja	LLaMA basiert, verlässlicher mit medizinischem Wissen
Vicuna [101]	ja	ja	fein-abgestimmter Chat-Bot basierend LLaMA
Alpaca [80]	ja	ja	fein-abgestimmtes LLaMA
LLaMA [84]	ja	nein	
Falcon [6]	ja	nein	keine
Guanaco-65B [26]	ja	ja	basiert auf LLaMA, besonders gut im multilingual Context
30B-Lazarus [18]	ja	nein	basiert auf LLaMA
BERT [41]	ja	nein	es gibt viele vortrainierten Modelle die verwendet werden können
Orca [58]	ja	ja	basiert auf LLaMA 2 mit synthetischen Daten
Phi-1 [38]	ja	nein	für Python code

LLM	vortrainiert	feinabgestimmt	Besonderheiten
StableLM [78]	ja	ja	trainiert mit verschiedenen Datensätzen auch für Code Intelligenz
FLAN-T5 [30]	ja	ja	von T5 trained on Flan-Collection
Stable-Beluga [77]	ja	ja	nutzt LLaMA trainiert auf synthetischen Daten
Dolly [25]	ja	ja	basiert auf pythia-12b
BloomZ [14]	ja	ja	ja unter BigScience Lizenz, Bloom(als multilingual language models)
OpenChatKit [24]	ja	nein	Sammlung an LLMs
BioGPT [50]	ja	nein	
BioMedLM [63]	ja	nein	in Zusammenarbeit mit MosaicLM
BioBERT [46]	ja	ja	
BioLinkBERT [96]	ja	nein	
ClinicalBERT [7]	ja	nein	
PMC-LLaMA [63]	ja	ja	
GatorTron [95]	ja	nein	
SciBERT [11]	ja	nein	

Tabelle A.2: Überblick über alle freien LLM

A.3 Datensätze und Benchmark

Aufgabe	Datensatz	BioBERT	SciBERT	BLUE	BLURB
NER	BC5-chem	ja	ja	ja	ja
	BC5-disease	ja	ja	ja	ja
	NCBI-disease	ja	ja		ja
	BC2GM	ja			ja
	JNLPBA	ja	ja		ja
	BC4CHEMD	ja			
	2010 i2b2/VA	ja			
	LINNAEUS	ja			
	Species-800	ja			
	SciERC		ja		
	ShARe/CLEF			ja	
PICO	EBM Pico				ja
	EBM-NLP		ja		
Extraktion von Beziehungen	ChemProt	ja	ja	ja	ja
	DDI			ja	ja
	GAD	ja			ja
	EU-ADR	ja			
	SciERC		ja		
	i2b2 2010			ja	
Satzähnlichkeit	BIOSSES			ja	ja
	MedSTS			ja	
Dokumentenklassifikation	HoC			ja	ja
Antworten auf Fragen	PubMedQA				ja
	BioASQ	ja			ja
Analyse von Abhängigkeiten	GENIA-LAS		ja		
	GENIA-UAS		ja		
Textklassifikation	ACL-ARC		ja		
	Paper Field		ja		
	SciCite		ja		
Schlussfolgerungsaufgabe	MedNLI			ja	

Tabelle A.3: Datensätze in den jeweiligen Benchmarks

Erklärung zur selbstständigen Bearbeitung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

Ort	Datum	Unterschrift im Original