Universität Leipzig
Medizinische Fakultät
Institut für Medizinische Informatik, Statistik und Epidemiologie

QUESTION ANSWERING AUF DEM LEHRBUCH 'HEALTH INFORMATION SYSTEMS' MIT HILFE VON UNÜBERWACHTEM TRAINING EINES PRETRAINED TRANSFORMERS

MASTERARBEIT

zur Erlangung des akademischen Grades

Master of Science
(M. Sc.)

vorgelegt von

Paul Keller
Studiengang Medizininformatik M. Sc.

Leipzig, den 31.09.2023

[23. Mai 2023 at 12:55 – classicthesis v4.6]

AUTOR:
Paul Keller
Geboren am 23.05.1998 in Leipzig, Deutschland

TITEL:
Question Answering auf dem Lehrbuch 'Health Information Systems' mit Hilfe von unüberwachtem Training eines Pretrained Transformers

INSTITUT:
Institut für Medizinische Informatik, Statistik und Epidemiologie
Medizinische Fakultät Universität Leipzig

REFERENT:
Prof. Dr. Alfred Winter

BETREUER:
Konrad Höffner

ABSTRAKT

DANKSAGUNG

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

AKRONYME

QAS	Question Answering System
GPT	General Pretrained Transformer
GMDS	Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V.
NN	Neuronales Netzwerk
SOTA	State of the Art (Stand der Technik)
LLaMa	Large Language Model Meta AI
LLM	Large Language Model

1 EINLEITUNG

1.1 GEGENSTAND

Eine effektive und effiziente Informationsbeschaffung bildet einen fundamentalen Bestandteil einer qualitativ hochwertigen klinischen Praxis in der Medizin. Bei jeder medizinischen Handlung werden große Mengen an Informationen genutzt und erzeugt, sei es als Grundlage für eine Diagnose oder zur Dokumentation des Behandlungsprozesses. Die strukturierte und klassifizierte Speicherung und Wiedergabe dieser Informationen ist ein kontinuierlicher Entwicklungsprozess und Gegenstand aktueller Forschung.

Die Digitalisierung der Medizin ist ein weites Themenfeld mit stetig wachsendem Bedarf. Die Medizinische Informatik beschreibt dabei „die Wissenschaft der systematischen Erschließung, Verwaltung, Aufbewahrung, Verarbeitung und Bereitstellung von Daten, Informationen und Wissen in der Medizin und im Gesundheitswesen. [...]“[1]. Vor diesem Hintergrund gewinnt die Entwicklung und Implementierung effizienter Informationssysteme und Technologien zur Unterstützung der klinischen Praxis zunehmend an Bedeutung.

In der Lehre wird die Praxis der Medizinischen Informatik durch umfangreiche Literatur, z.B. in Winter u. a. (2023), unterstützt. Zur Strukturierung von Fachbegriffen und Rollen des Informationsmanagements im Krankenhaus existiert die Ontologie SNIK (Jahn u. a., 2014), ein semantisches Netz, kategorisiert in der Metaontologie SNIK und Teil des Projekts SNIK des Instituts für Medizinische Informatik, Statistik und Epidemiologie[2] an der Universität Leipzig. Die Nutzung dieses Netzes ermöglicht eine systematische Darstellung von Rollen, Entitäten und Funktionen des Informationsmanagements im Krankenhaus, unabhängig von der Definition der zugrunde liegenden Literaturquellen.

Die Bedeutung von maschinellem Lernen, Deep Learning und Sprachmodellen ist in der heutigen Zeit sehr präsent. Diese Technologien werden in vielen Bereichen, von der Automobilindustrie bis hin zu medi-

1 Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS) (2023). Definition Medizinische Informatik. https://www.gmds.de/aktivitaeten/medizinische-informatik/ (abgerufen am 07. 03. 2023).

2 Institut für Medizinische Informatik, Statistik und Epidemiologie. https://www.imise.uni-leipzig.de/Institut (besucht am 09. 03. 2023).

zinischen Anwendungen, eingesetzt, um neue Methoden der Informationsgewinnung und -verarbeitung zu ermöglichen. Verschiedene KI-Modelle können in vielen Bereichen, wie z.B. der Erkennung seltener Krankheiten (Brasil u. a., 2019), der personalisierten Medizin (Johnson u. a., 2021) oder dem Marketing (Davenport u. a., 2020), weitreichende Auswirkungen auf zukünftige Arbeitsprozesse haben. Sprachmodelle wie das General Pretrained Transformer (GPT)-3-Modell von OpenAI (Brown u. a., 2020) können dazu beitragen, Texte in verschiedenen Sprachen automatisch zu übersetzen und sogar kreatives Schreiben zu ermöglichen. Deep Learning, das auf künstlichen neuronalen Netzen basiert, ermöglicht eine noch tiefere und komplexere Verarbeitung von Daten. In der Medizin kann Deep Learning beispielsweise zur Diagnose von Krankheiten und zur Analyse medizinischer Bilder eingesetzt werden (Esteva, Kuprel, Novoa u. a., 2017).

1.2 PROBLEMSTELLUNG

Informationssysteme im Gesundheitswesen sind hochkomplex, abhängig von den Anforderungen der jeweiligen Organisation und unterliegen einer ständigen Weiterentwicklung. Die Anforderungen sind vielfältig und umfassen beispielsweise die Speicherung und Verarbeitung von Patientendaten, die Dokumentation von Behandlungsprozessen, die Unterstützung von Diagnose- und Therapieprozessen sowie die Unterstützung von Forschungsaktivitäten. Ausgehend von diesen Anforderungen gibt es eine Fülle von Literatur, die sich mit der Definition, Implementierung und Wartung von Informationssystemen im Gesundheitswesen befasst. Das Buch *Health Information Systems* von Winter u. a. beschäftigt sich umfassend mit diesen Anforderungen und ist im März 2023 in der 3. Auflage erschienen (Winter u. a., 2023).

Das Management von Informationssystemen ist eine anspruchsvolle Aufgabe, die sich nicht nur auf die Anwendung durch das Krankenhauspersonal beschränkt. Es ist auch von großer Bedeutung für Studierende, um ein besseres Verständnis der bestehenden Systeme zu erlangen, und für Wissenschaftler, um diese Systeme zu erweitern oder neue Managementmethoden zu entwerfen. Eine konkrete und konsistente Wissensbasis ist sowohl für die Anwendung als auch für Lehre und Forschung von großer Bedeutung.

Eine weitere Herausforderung bei der Auseinandersetzung mit der Literatur zu Informationssystemen ist die Komplexität der Übertragung von theoretischen Konzepten auf praktische Anwendungsfälle. Insbesondere für Studierende und Praktiker erfordert die praktische Anwendung ein tiefes Verständnis der Zusammenhänge und der Anwendbarkeit der vorgestellten Konzepte auf konkrete Arbeitsumge-

bungen. Da die verfügbare Literatur oft sehr umfangreich und ihre Definitionen fragmentiert sind, stellt die Identifikation relevanter Informationen für spezifische Problemstellungen eine weitere Herausforderung dar. Die Fragmentierung von Definitionen bezieht sich hier auf die Erläuterung eines Fachbegriffs oder Konzepts innerhalb einer Literaturquelle. Diese Definitionen werden häufig nicht zusammenfassend aufgelistet, sondern ergeben sich im Laufe der Texte und Kapitel. Dies führt dazu, dass große Teile eines Buches gelesen werden müssen, um einzelne Konzepte und ihre Beziehungen zu anderen Themen vollständig zu erfassen.

Der Umfang der Literaturquellen und die Fragmentierung bei der Definition von Fachbegriffen erschweren eine schnelle Informationsbeschaffung, insbesondere für Studierende, die grundlegende Konzepte richtig verstehen wollen.

- Problem: Schwierigkeiten bei der Informationsbeschaffung aufgrund des Umfangs von Winter u. a. (2023) und der Fragmentierung von Definitionen

1.3 MOTIVATION

Eine Strukturierung von Informationen zum Management von Krankenhausinformationssystemen ist bereits Bestandteil des Projekts SNIK (Jahn u. a., 2014; Schaaf u. a., 2015), ein semantisches Netz des Instituts für Medizinische Informatik, Statistik und Epidemiologie der Universität Leipzig. Auf Basis dieses Netzes wurden bereits verschiedene Methoden zur Informationsextraktion untersucht.

Die BeLL mit dem Titel „Question Answering on SNIK“ (Brunsch, 2022) erweiterte den Zugang zu diesem Netz durch die Verwendung natürlicher englischer Sprache. Die Ergebnisse der BeLL wurden in einem Question Answering System (QAS) QAnswer (QAnswer, 2023) umgesetzt, zeigen aber Defizite in der Erklärbarkeit und Verständlichkeit komplexerer Fragen. Im Gegensatz dazu untersuchte Omar u. a. (2023) auf einem anderen Datensatz die Leistung einer Konversations-KI (in diesem Fall ChatGPT) im Vergleich zu einer herkömmlichen, auf einem Wissensgraphen basierenden QAS (verwendet wurde $KGQA_N$). Die Ergebnisse zeigten, dass ChatGPT im Vergleich zum verwendeten $KGQA_N$ erstaunlich stabile und verständliche Antworten lieferte, jedoch bei der Ausgabe korrekter Antworten deutlich schlechter abschnitt. Hier steht zu erwarten, dass durch ein besseres Feintuning die Anzahl der falschen Antworten deutlich reduziert werden kann.

Es ist daher notwendig zu untersuchen, ob der Einsatz einer Konversations-KI die Anwendungsschwierigkeiten in QAnswer beheben kann und

ob die Wiedergabe mit niedrigem Wahrheitswert durch Feintuning verbessert werden kann.

1.4 ZIELSETZUNG

Dem in 1.2 gezeigten Problem werden folgende Ziele dieser Arbeit zugeordnet.

- Ziel Z1: Beantwortung von Fragen zu Informationssystemen im Gesundheitswesen in natürlicher Sprache durch eine Konversation-KI mit Hilfe von Winter u. a. (2023)
- Ziel Z2: Lösung einer Beispielklausur des Moduls „Architektur von Informationssystemen im Gesundheitswesen“[3] mit Hilfe einer Konversations-KI. Das Ziel ist kein produktives System, sondern lediglich die Machbarkeit der Beantwortung von Fragen mit Hilfe einer Konversations-KI aufzeigen.

1.5 BEZUG ZU ETHISCHEN LEITLINIEN DER GMDS

Die ethischen Leitlinien der GMDS (Ahrens u. a., 2023) geben „sowohl den tragenenden Gesellschaften als Institution als auch dem einzelnen Mitglied eine Orientierung, welche ethischen Forderungen in ihrem bzw. seinem jeweiligen Aufgaben- und Verantwortungsbereicht relevant sein können“ (Ahrens u. a., 2023). Aufgeteilt in 16 Artikel werden hier verschiedene Kompetenzen und Verantwortlichkeiten definiert, unter die auch das hier beschriebene System fällt. Da die zu entwickelnde Konversations-KI Informationen über medizinisches Wissen, insbesondere über Informationssysteme im Gesundheitswesen, bereitstellt, unterliegt sie in ihrer Existenz als „Informationsbasis“ einer Vielzahl von Artikeln. Sie wirkt unterstützend in den Artikeln 1 „Auftrag“, 2 „Fachkompetenz“, 3 „Kommunikative Kompetenz“, kann aber durch unsachgemäßen Gebrauch und ihr zugrundeliegendes Wesen entgegen den Artikeln 1 „Auftrag“, 2 „Fachkompetenz“, 3 „Kommunikative Kompetenz“, 4 „Medizinethische Kompetenz“, 6 „Soziale Verantwortung“ und 13 „Forschung“ handeln.

In Artikel 1 „Auftrag“ heißt es unter anderem, dass „Die Würde des Menschen und das Persönlichkeitsrecht [...] dabei vorrangig geachtet und geschützt werden [müssen]“ (Ahrens u. a., 2023). Dazu gehört insbesondere die „Allgemeine Erklärung der Menschenrechte“, in der das „Verbot der Diskriminierung z.B. nach Geschlecht oder [aus rassistischen Gründen] (Art. 2, 7)“ verankert ist und der „Schutz des (geistigen) Eigentums (Art. 17, 27)“ und die „Gedanken-, Gewissens-, Religions- und Meinungsfreiheit (Art. 18, 19)“ beschrieben werden.

3 einem Modul des Masterstudiengangs Medizinische Informatik an der Universität Leipzig, das inhaltlich auf Winter u. a. (2023) aufbaut

Die von der Konversations-KI zu generierenden Antworten ergeben sich aus einer Vielzahl von zuvor genutzten Datensätzen aus dem Internet, die nicht notwendigerweise diesen ethischen Leitlinien folgen. Eine inhaltliche Garantie für die Einhaltung dieser Leitlinien ist daher ohne zusätzliche Filterung der Antworten zu gestellten Fragen nicht möglich. Durch eine optimale Filterung können Antworten, die den Leitlinien widersprechen, ausgeschlossen werden. Diese Filterung ist jedoch aufgrund ihrer Komplexität sowohl in zeitlicher als auch in finanzieller Hinsicht nicht Bestandteil dieser Arbeit. Es muss daher davon ausgegangen werden, dass es durchaus Antworten geben kann, die gegen die Leitlinien verstoßen.

Artikel 1 wird durch die Konversations-KI gefördert, in dem sie Medizinische Versorgungseinrichtungen durch effiziente und schnelle Informationsbeschaffung in ihren Fähigkeiten unterstützt, „ihre Leistungen qualitativ und quantitativ nachweisen, überwachen und sicherstellen [zu] können“ (Ahrens u. a., 2023).

Artikel 2 „Fachkompetenz“ definiert, dass das Mitglied seine „Fachkompetenz nach dem Stand der Wissenschaft und Technik erwirbt [. . .]“ und „Maßnahmen zur Fehlervermeidung ergreift“ (Ahrens u. a., 2023). Dies ist teilweise durch die Konversations-KI gegeben, da die extrahierten Informationen auf dem Stand des zugrundeliegenden Buches Winter u. a. (2023) sind. Eine Fehlervermeidung ist hier jedoch nicht vorgesehen. Wie in Omar u. a. (2023) gezeigt, weist ChatGPT als GPT-Modell eine geringe Wahrheitsquote auf. Da es sich bei der zu entwickelnden Konversations-KI um ein GPT-Modell handelt, wird die fehlerfreie Beantwortung von Fragen zwar angestrebt, kann aber nicht garantiert werden.

Artikel 3 „Kommunikative Kompetenz“ beschreibt die Fähigkeit „Recht, Interessen [und] Konventionen der verschiedenen von [dem Mitglied] seiner Arbeit Betroffenen zu verstehen und zu berücksichtigen [. . .]“ und „Wissenschaftliche Erkentnisse in verständlicher Form der Öffentlichkeit zugänglich [. . .]“ (Ahrens u. a., 2023) zu machen. Die Konversations-KI unterstützt dabei Fragende durch eine Antwort in möglicherweise verständlicherer Form der wissenschaftlichen Erkenntnisse in Winter u. a. (2023), jedoch ist hier nicht gegeben, dass die Betroffenen in der Antwort berücksichtigt oder verstanden wurden.

Artikel 4 „Medizinethische Kompetenz“ legt fest, dass das Mitglied „ethische Prinzipien der Medizin [. . .] bei seinem beruflichen Handeln beachtet“ (Ahrens u. a., 2023). Zu den ethischen Prinzipien (Ahrens u. a., 2023) gehören auch die „Achtung vor dem Menschen zu wahren“ und die „Würde des Invidiums zu schützen“. Beides kann nicht

allein durch die Konversations-KI gewährleistet werden und muss durch mögliche Filter ergänzt werden.

Artikel 6 „Soziale Verantwortung“ definiert die Verantwortungen des Mitglied „gesellschaftliche Auswirkungen [zu] berücksichtigen [...]“ und die „Allgemeine Erklärung der Menschenrechte und ethische[n] Prinzipien der Medizin“ (Ahrens u. a., 2023)zu beachten. Wie bereits zu den Artikeln 1 und 4 ausgeführt, ist es der Konversations-KI aufgrund der Datengrundlage nicht möglich, diesen Artikel, insbesondere die hier genannten Punkte, in ihrer Antwort zu berücksichtigen.

Artikel 13 „Forschung“ definiert, dass das Mitglied „gute Wissenschaftliche Arbeit, inbesondere Offenheit und Transparenz [und] Akzeptanz von Kritik“ (Ahrens u. a., 2023) einhalten soll. Gute wissenschaftliche Arbeit wird weiter definiert als „die strikte Ehrlichkeit im Hinblick auf die Beiträge von Partnern, Konkurenten, Vorgängern zu wahren“ und „weder Fälschung oder Plagiate [zu] benutzen“ (Ahrens u. a., 2023). Da die Konversations-KI auf dem Inhalt eines Buches trainiert wird, in ihrer Datenbasis aber bereits eine Vielzahl von Texten, darunter auch Plagiate und Fälschungen, gelernt hat, ist es nicht möglich, die Antworten als wissenschaftliche Quelle zu verwenden, da sowohl Plagiate als auch Fälschungen ohne Annotation vorhanden sein können. Die Antworten sollten als reine Informationsextraktion und nicht als Quelle für wissenschaftliche Arbeiten betrachtet werden.

1.6 AUFGABENSTELLUNG

Die in 1.4 genannten Ziele Z_i werden durch die hier aufgeführten Aufgaben A_i gelöst.

- Aufgabe zu Ziel Z1
  - Aufgabe A1.1: Es sollen aktuelle Sprachmodelle verglichen werden. Dabei sind die Einschränkungen der Verfügbarkeit und Verwendbarkeit zu berücksichtigen. GPT-Modelle basieren auf großen Datenmengen und enthalten mehrere Milliarden Parameter. Ein eigenes Training eines Modells würde sowohl den zeitlichen als auch den finanziellen Rahmen übersteigen, weshalb auf ein vortrainiertes Modell zurückgegriffen werden muss. Dazu muss das Modell frei verfügbar sein und unter einer Open-Source-Lizenz stehen. Außerdem muss das Modell am Rechenzentrum der Universität Leipzig geladen und trainiert werden können. Aufgrund der großen Anzahl an Parametern sind auch hier Grenzen des Arbeitsspeichers und damit der Größe des Modells gesetzt.

  - Aufgabe A1.2: Für ein effizientes und erfolgreiches Feintuning der Konversations-KI ist eine Datenkuration von Winter u. a. (2023) notwendig. Abschnitte wie das Literaturverzeichnis, Aufzählungen oder Grafiken und deren Beschriftung müssen vor der Verwendung des Textes entfernt oder umgeschrieben werden.
  - Aufgabe A1.3: Die trainierte Konversations-KI wird dann zur Beantwortung von Fragen zu Winter u. a. (2023) verwendet. Dabei wird während des Trainings der in Brown u. a. (2020) beschriebene „Zero-Shot"-Ansatz verfolgt. Dies bedeutet, dass das Verständnis der Fragestellung und die Bewertung wichtiger Informationen allein durch das Modell erfolgt und nicht durch vordefinierte Fragen im Datensatz dem Modell beigebracht wird.

- Aufgabe zu Ziel Z2
  - Aufgabe A2.1: Die in dieser Arbeit erstellte Konversations-KI wird vor und nach dem Training hinsichtlich ihrer Fähigkeit, Fragen korrekt zu beantworten, evaluiert. Dies ermöglicht eine Aussage über die Effektivität des Trainings und die Leistungssteigerung der Konversations-KI. Ebenso wird ein Vergleich mit dem aktuellen GPT-4 Modell (OpenAI, 2023) durchgeführt, um die Notwendigkeit eines Feintunings zu ermitteln.
  - Aufgabe A2.2: Bewertung der Antwortoptionen von Klausurfragen nach gleichen Kriterien wie in Omar u. a. (2023)

1.7 AUFBAU DER ARBEIT

Kapitel 1 beschreibt das grundlegende Umfeld dieser Arbeit, formuliert exitierende Probleme und Anforderungen, bietet Ziele zur Lösung dieser Probleme an und gibt Aufgaben, zur Umsetzung dieser Ziele. In Kapitel 2 werden Grundlagen gelegt zum Verständnis der in dieser Arbeit verwendeten Technologie, während in Kapitel 3 der aktuelle Stand der Forschung zusammengefasst wird. Kapitel 4 umfasst Lösungsstrategien des in 1.2 formulierten Problem mit einer anschließenden Beschreibung der Umsetzung dieser Lösungen in Kapitel 5. Die Ergebnisse dieser Arbeit werden in Kapitel 6 präsentiert und in Kapitel 7 zusammengefasst diskutiert. Zusätzlich gibt Kapitel 7 einen Ausblick dieser Arbeit.

2

GRUNDLAGEN

Themen: - General Pretrained Transformer - Neuronal Networks - Zero Shot Ansatz - Finetuning - Datenkuration

Topics: - GPT als Chat - Wissenextraktion mit GPT - Finetuning Variables

3 STAND DER FORSCHUNG

3.1 GRUNDLAGEN DER ARCHITEKTUR

Die erste schriftliche Erwähnung des Transformer-Models und zusätzlich auch Einführung der zwei Teil-Modelle Encoder und Decoder findet sich in Vaswani u. a. (2017). Die hier beschriebene bidirektionale Architektur beinhaltet jedoch die Grundlage aller darauf aufbauenden Modelle und Weiterentwicklungen. Die grundlegene Architektur wurde für unterschiedliche Anwendungen stark modifiziert. Seit 2017 gibt es grundsätzliche Differenzen in den Modellen und deren Möglichkeiten. Deshalb führte Kalyan, Rajasekharan und Sangeetha (2022) eine Taxonomie der Transformer-basierenden Vortrainierten Sprachmodelle. Dieser Taxonomie wird hier zur Beschreibung weitere Architekturen und Methodiken gefolgt.

3.2 WEITERENTWICKLUNG DER ARCHITEKTUR

Neben dem Grundbaustein eines Transformers - dem Attention-Neuronales Netzwerk (NN) - sind zwei wichtige Änderungen zu normalen neuronalen Netzwerken in Transformer eingeflossen. Restverbindungen als Ebenen-Normalizierung, im Englischen „Deep Residual Connections“, verändern die Zielsetzung eines NN, behalten jedoch durch ihre Ebenen-Normalisierung die selben Ausgaben. In He u. a. (2016) wurde dieses Konzept erstmals eingeführt und liefert die Lösung zu einem Grundproblem von großen, aus mehreren Ebenen bestehenden Transformer-Modellen. Es zeigte sich schon 2016 im Bereich der Bilderkennung, das mit steigender Tiefe die Korrektheit von Modellen sich sätigt und dann schnell verschlechtert, sollte jenes Modell weitertrainiert werden. Dies setzte ein praktische Grenze der Tiefe von NNs und verhinderte somit komplexere Probleme mit größeren Modellen zu lösen. He u. a. (2016) beschreiben eine Lösung durch die genannten Restverbindungen, welche normale NN simple ersetzen können, und belegen ebenso die Effektivität dieser Methode.

Die zweite wichtige Änderung ist die Einführung von Dropout. Dropout ist eine Methode, welche die Trainingszeit von NNs verkürzt und die Generalisierung verbessert. Srivastava u. a. (2014) beschreiben die Methode als das zufällige Aussetzen von Neuronen in einem NN. Dieses Aussetzen wird zufällig gewählt und ist nicht von der Eingabe abhängig. Durch das Aussetzen von Neuronen wird das NN

gezwungen, sich nicht auf andere Neuronen zu verlassen und somit eine bessere Generalisierung zu erreichen. Dieses Aussetzen wird nur während des Trainings durchgeführt und nicht während der Inferenz. Die Methode wurde 2014 eingeführt und ist seitdem ein fester Bestandteil von NNs.

3.3 CONTINUAL PRETRAINING UND DIE NUTZUNG VON SPRACHMODELLEN

Ein Tranformer-Modell als Wissensbasis ist in Omar u. a. (2023) verglichen mit verschiedenen State of the Art (Stand der Technik) (SOTA)-Modellen. Sie zeigen eine deutliche Verbesserung der Robustheit gegenüber fehlerhafter Eingabe, Erklärbarkeit von Antworten und Fragenverständnis von komplexeren Fragen mit mehreren Fakten durch ChatGPT, zeigen allerdings Probleme in der Aktualität von Informationen, dem Wissen zu spezifischen Domänen und dem wohl wichtigsten, der korrekten Beantwortung von Fragen. Grund hierfür ist eine einerseits grundlegene Eigenschaft von GPT-Modellen, keine Inkorperation von aktuellen Informationen. Das Trainieren von GPT-Modellen ist ein aufwändiger Prozess und kann nicht bei jeder Inferenz (der Nutzung des Modells durch die Generierung von Text) durchgeführt werden. Desweiteren wurde ChatGPT jedoch ohne zusätzliches Continual Pretraining genutzt. Eine Anpassung auf Domänen und Verbesserung der Korrektheit von Antworten steht somit noch aus.

Radford und Narasimhan (2018) zeigen die Verbesserung der Leistung von Transformer-Modellen durch Generative Pretraining und eine weitere Steigerung dieser durch überwachtes Fine-Tuning. Auch hier zeigt sich eine deutlicher Trend. Mit steigender Größe des Datensatzes, steigender Länge des Trainingprozesses und steigender Größe der Modelle verbessern sich die Ergebnisse von Modellen.

Diesen Trend belegen Kaplan u. a. (2020) und berechnen hier den Einfluss von verschiedenen Einflussgrößen auf die Gesamtleistung eines Modells. Durch die hier genutzen Einflussgrößen lässt sich eine Vorhersage der Leistung eines Modells treffen. Der Artikel endet mit einer Vermutung auf die theoretische maximale Leistung und damit maximale Größe von Transformer-Modellen.

Um diesen beschriebenen Skalierungsregeln zu folgen, jedoch die Trainigszeit und notwendige Datenmenge zu reduzieren, gibt es die Möglichkeit Continual Pretraining zu nutzen. Gururangan u. a. (2020) wendeten diese Methodik an und zeigten, dass Modelle immens davon profitieren, Domäne-spezifisches Wissen zu adaptieren und aus der großen Menge an grundlegenden Daten bessere korrekte Ant-

worten in einer spezifischen Domäne zu generieren. Erstmals in Lee u. a. (2019) genutzt um das Basismodell BERT auf die biomedizinische Domäne anzupassen, erweitern Gururangan u. a. (2020) diese Methode und zeigen die Anwendbarkeit auf verschiedene Domänen und Aufgaben. Das Continual Pretraining auf Aufgaben-Spezifische Daten verbessert die Leistung für spezifischen Aufgaben, während die Trainingszeit 60 mal kürzer ausfällt im Vergleich zu Continual Pretraining auf Domäne. Eine Verbindung beider Arten liefert hier die besten Ergebnisse.

Doch nicht nur Continual Pretraining verbessern die Ausgaben der Modelle, sondern auch das überwachtes Fine-Tuning. Ziegler u. a. (2019) beschreiben in ihrem Artikel die Effektivität von Reinforcement Learning als Fine-Tuning Methode um die Aufgaben der Weiterführung und Zusammenfassung von Texten zu lösen. Fine-Tuning benötigt jedoch gekennzeichnete Daten (engl. „labeled data“, Daten mit bekannten korrekten Ausgaben), welche in der Regel aufwändig zu Erstellen sind und nicht immer in der notwendigen Menge zur Verfügung stehen. In dieser Arbeit wird von einem Fine-Tuning durch die fehlende Verfügbarkeit von gekennzeichneten Daten abgesehen.

3.4 AKTUELLE MODELLE UND DEREN NUTZBARKEIT

Mit der Feststellung, dass die Leistung von Modellen mit steigender Größe, Trainigszeit und Daten steigt, wurden eine Reihe an Modellen entworfen, welche unterschiedlichste Architekturen, Anwendungsfälle und Leistungen besitzen. OpenAI erreichten mit GPT-3 einen Durchbruch in Popularität und beschrieben ihr Vorgehen in Brown u. a. (2020). In ihrem Artikel belegten sie die Leistungssteigerung durch größere Modelle und zeigten, dass diese Leistung ebenso ohne Fine-Tuning erreicht werden kann. Auch verglichen sie das Verhalten der Antworten abhängig von FewShot und ZeroShot Eingaben, wobei ersteres bessere Ergebnisse erzielten. Diese Erkenntnisse unterstützen die Annahme, dass auch ohne Fine-Tuning das in dieser Arbeit verwendete Modell gute Leistungen erreichen kann.

Weiterführend im Jahr 2023 veröffentlichte OpenAI GPT-4 und stellten dieses in OpenAI (2023) vor. Neben den weit aus besseren Ergebnissen durch ein noch größeres Modell mit mehr Parametern gelang es ihnen, nun auch Bild-Daten als Eingabe zu verarbeiten. Dieser Artikel wiederum unterstreicht die Annahme, dass größere Modelle bessere Leistung bringen und ein besseres Verständnis der natürlichen Sprache besitzen. Eine Nutzung dieses Modells, ebenso wie GPT-3 ist nicht möglich, da diese Modelle zum aktuellen Zeitpunkt nicht veröffent-

licht wurden.

In Kontrast zu den bisherigen Modellen veröffentlichten Black u. a. (2022) GPT-NeoX. Ein Modell, welches in seiner Größe und Leistung GPT-3 ähnelt, jedoch auf der Architektur von GPT-J basierend im Open-Source Rahmen veröffentlicht wurde. Sie zeigten, dass die meisten interessanten Fähigkeiten eines Modells erst ab eine bestimmen Anzahl an Parametern gezeigt werden.

Zuletzt veröffentlichte Touvron u. a. (2023) die Large Language Model Meta AI (LLaMa)-Modelle in unterschiedlicher Größe. Ein klarer Vorteil gegenüber anderen Modellen in ihrer Nutzbarkeit ist hier der Fokus auf längere Trainings-Zeit und einem größeren Datensatz gegenüber der Größe des Modells. Sie zeigten bessere Ergebnisse in fast allen Aufgabenbereichen gegenüber anderen Modellen wie GPT-3 und **PaLM!** (**PaLM!**) mit wesentlich weniger Parametern. Dadurch ist eine Nutzung jener Modelle billiger und schneller, einfacher und schneller zu trainieren mit gleichen oder besseren Ergebnissen. Auch diese Modelle wurden veröffentlicht und stehen somit zur Auswahl für diese Arbeit.

3.5 FORSCHUNG UND PROBLEME VON MODELLEN

Neben der Entwicklung von neuen Modellen wurden auch neue Ansätze zur besserem Continual Pretraining und Adaption von Modellen entworfen. Pfeiffer u. a. (2020) stellten in ihrem Artikel den Adapter vor, welche ein Einsatz von zusätzlichen NNs in verschiedene Ebenen der Transformer-Architektur ermöglichen. Durch Sie lässt sich die Adaption zu anderen Aufgaben und Domänen ohne Continual Pretraining des gesamten Modells erreichen, da während des Trainings sämtliche Parameter des Ursprungmodells fixiert bleiben, während die neu eingefügten Adapter trainiert werden. Zusätzlich lassen sich dadurch bereits vortrainierte Adapter zu weiteren Domänen und Aufgaben in aktuelle Modelle einfügen, ohne die Notwendigkeit jeglichen Trainings.

Dai u. a. (2022) untersuchten die Eigenschaften von Large Language Model (LLM)s auf ihre Eigenschaft, faktisches Wissen wiedergeben zu können, ohne eine Wissensdatenbank als Grundlage während der Nutzung zu besitzen. Sie stellten fest, dass besonders in weiter hinten liegenden Ebenen die Neuronalen Netze sogenannte „Wissensneuronen“ besitzen, die zu bestimmen Fakten korrelieren. Diese Wissensneuronen aktiveren sich, wenn ein bestimmter Fakt in der Eingabe angesprochen wird und können mittels Verstärkung oder Unterdrückung dazu führen, dass das Modell diesen Fakt besser berücksichtigt oder

„vergisst“.

Neben den überaus großen Erfolgen von neuen Modellen erheben sich jedoch auch neue Probleme bei der Benutzung dieser Modelle. Neben Falschaussagen ergeben sich Probleme durch sozialen und anderen Bias in den Antworten, Selbstüberschätzung bei falschen Aussagen, welches wiederum zu schwerwiegenden Problemen in der Anwendung dieser Modelle kommen kann, Generierung von schädlichen Inhalten, Unterstützung von Kriminalität mit Expertise und weiteren Probleme. OpenAI (2023) dedizierten einen eigenen Abschnitt ihres Artikels zur Untersuchung dieser Probleme und deren Adressierung. Sie zeigten hier grundlegende Probleme bei der Anwendung von Sprachmodellen auf, hielten jedoch konkreten Lösungsansätzen zurück.

In Dehouche (2021) beschreibt der Author weitere Fragestellungen zu dem Umgang mit Antworten von Sprachmodellen und deren Konflikt zum Urheberrecht. Wem gehört der generierte Text - den Authoren der Datensätze auf dem das Modell trainiert wurde, der Firma dem das Modell gehört, dem Nutzer der das Modell anleitet? Auch hier zeigen sich ungelöste Probleme in der Anwendung von Sprachmodellen und bieten Raum für weitere Forschung.

4

LÖSUNGSANSATZ

5 AUSFÜHRUNG DER LÖSUNG

6 ERGEBNISSE

7 DISKUSSION

ZUSAMMENFASSUNG

LITERATUR

Ahrens, Wolfgang u. a. (2023). *Ethische Leitlininen der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS), des Arbeitskreises der IT-Leiter/innen der Universitätsklinika (AL-KRZ) des Berufsverbandes Medizinischer Informatiker (BVMI), des Bundesverbandes der Krankenhaus-IT-Leiterinnen/Leiter e.V. (KH-IT) und des Deutschen Verbandes Medizinischer Dokumentare e.V. (DVMD)*. URL: https://www.gmds.de/fileadmin/user_upload/Aktivitaeten_Themen/praesidiumskommissionen/Ethische_Leitlinien.pdf (besucht am 23. 04. 2023).

Black, Sidney u. a. (Mai 2022). „GPT-NeoX-20B: An Open-Source Autoregressive Language Model“. In: *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics, S. 95–136. DOI: 10.18653/v1/2022.bigscience-1.9.

Brasil, Sandra, Carlota Pascoal, Rita Francisco, Vanessa dos Reis Ferreira, Paula A. Videira und Gonçalo Valadão (2019). „Artificial Intelligence (AI) in Rare Diseases: Is the Future Brighter?“ In: *Genes* 10.12. ISSN: 2073-4425. DOI: 10.3390/genes10120978.

Brown, Tom u. a. (2020). „Language Models are Few-Shot Learners“. In: *Advances in Neural Information Processing Systems*. Hrsg. von H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan und H. Lin. Bd. 33. Curran Associates, Inc., S. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Brunsch, Hannes Raphael (2022). „Question Answering auf SNIK“. Besondere Lernleistung. Leipzig, Germany: Wilhelm-Ostwald-Schule. URL: https://www.snik.eu/public/bell-hrb.pdf.

Dai, Damai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang und Furu Wei (2022). *Knowledge Neurons in Pretrained Transformers*. arXiv: 2104.08696 [cs.CL].

Davenport, Thomas, Abhijit Guha, Dhruv Grewal und Timna Bressgott (2020). „How artificial intelligence will change the future of marketing“. In: *Journal of the Academy of Marketing Science* 48.1, S. 24–42. ISSN: 1552-7824. DOI: 10.1007/s11747-019-00696-0.

Dehouche, Nassim (März 2021). „Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3): “The best time to act was yesterday. The next best time is now.”“ In: *Ethics in Science and Environmental Politics* 21. DOI: 10.3354/esep00195.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa u. a. (2017). „Dermatologist-level classification of skin cancer with deep neural networks“. In: *Nature* 542, S. 115–118. DOI: 10.1038/nature21056.

Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey und Noah A. Smith (Juli 2020). „Don't Stop Pretraining: Adapt Language Models to Domains and Tasks". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, S. 8342–8360. DOI: 10.18653/v1/2020.acl-main.740. URL: https://aclanthology.org/2020.acl-main.740.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren und Jian Sun (2016). „Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 770–778. DOI: 10.1109/CVPR.2016.90.

Jahn, Franziska, Michael Schaaf, Barbara Paech und Alfred Winter (2014). „Ein Semantisches Netz des Informationsmanagements im Krankenhaus". In: *Informatik 2014*. Hrsg. von E. Plödereder, L. Grunske, E. Schneider und D. Ull. Lecture Notes in Informatics. Bonn: Gesellschaft für Informatik e.V., S. 1491–1498.

Johnson, Kevin B., Wei-Qi Wei, Dilhan Weeraratne, Mark E. Frisse, Karl Misulis, Kyu Rhee, Juan Zhao und Jane L. Snowdon (2021). „Precision Medicine, AI, and the Future of Personalized Health Care". In: *Clinical and Translational Science* 14.1, S. 86–93. DOI: 10.1111/cts.12884.

Kalyan, Katikapalli Subramanyam, Ajit Rajasekharan und Sivanesan Sangeetha (2022). „AMMU: A survey of transformer-based biomedical pretrained language models". In: *Journal of biomedical informatics* 126, S. 103982. DOI: 10.1016/j.jbi.2021.103982.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu und Dario Amodei (2020). *Scaling Laws for Neural Language Models*. arXiv: 2001.08361 [cs.LG].

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So und Jaewoo Kang (2019). „BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4, S. 1234–1240. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz682.

Omar, Reham, Omij Mangukiya, Panos Kalnis und Essam Mansour (2023). *ChatGPT versus Traditional Question Answering for Knowledge Graphs: Current Status and Future Directions Towards Knowledge Graph Chatbots*. Version 1. arXiv: 2302.06466 [cs.CL].

OpenAI (2023). „GPT-4 Technical Report". In: arXiv: 2303.08774 [cs.CL].

Pfeiffer, Jonas, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho und Iryna Gurevych (Okt. 2020). „AdapterHub: A Framework for Adapting Transformers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, S. 46–54. DOI: 10.18653/v1/2020.

emnlp-demos.7. URL: https://aclanthology.org/2020.emnlp-demos.7.

QAnswer (2023). *Question Answering*. URL: https://app.qanswer.ai/public-share?kb=SNIK_BB&type=graph&user=kirdie (besucht am 09. 03. 2023).

Radford, Alec und Karthik Narasimhan (2018). „Improving Language Understanding by Generative Pre-Training". In.

Schaaf, Michael, Franziska Jahn, Kais Tahar, Christian Kucherer, Alfred Winter und Barbara Paech (2015). „Entwicklung und Einsatz einer Domänenontologie des Informationsmanagements im Krankenhaus". In: *Informatik 2015*. Lecture Notes in Informatics 246. Hrsg. von Douglas W. Cunningham, Petra Hofstedt, Klaus Meer und Ingo Schmitt.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever und Ruslan Salakhutdinov (2014). „Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.56, S. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

Touvron, Hugo u. a. (2023). *LLaMA: Open and Efficient Foundation Language Models*. arXiv: 2302.13971 [cs.CL].

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser und Illia Polosukhin (2017). „Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Hrsg. von I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan und R. Garnett. Bd. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Winter, Alfred, Elske Ammenwerth, Reinhold Haux, Michael Marschollek, Bianca Steiner und Franziska Jahn (2023). *Health Information Systems*. 3. Aufl. Health Informatics. Springer Cham. ISBN: 978-3-031-12310-8. DOI: 10.1007/978-3-031-12310-8. URL: https://link.springer.com/book/10.1007/978-3-031-12310-8.

Ziegler, Daniel M., Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano und Geoffrey Irving (2019). „Fine-Tuning Language Models from Human Preferences". In: *CoRR* abs/1909.08593. URL: http://arxiv.org/abs/1909.08593.

APPENDIX

ERKLÄRUNG

Ich versichere, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet.
Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann.
Ich versichere, dass das elektronische Exemplar mit den gedruckten Exemplaren übereinstimmt.

Leipzig, 31.09.2023

Paul Keller