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

[14. Juni 2023 at 15:32 – 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

Abstrakt iii

1 Einleitung 1
- 1.1 Gegenstand . . . 1
- 1.2 Problemstellung . . . 2
- 1.3 Motivation . . . 3
- 1.4 Zielsetzung . . . 4
- 1.5 Bezug zu ethischen Leitlinien der GMDS . . . 4
- 1.6 Aufgabenstellung . . . 6
- 1.7 Aufbau der Arbeit . . . 7

2 Grundlagen 9
- 2.1 Transformer . . . 9
- 2.2 Tokenization . . . 10
  - 2.2.1 Byte-Pair-Encoding . . . 11

3 Stand der Forschung 13
- 3.1 Continual Pretraining und die Nutzung von Sprachmodellen . . . 13
- 3.2 Aktuelle Modelle und deren Nutzbarkeit . . . 14
- 3.3 Forschung und Probleme von Modellen . . . 15

4 Lösungsansatz 19
- 4.1 Auswahl von Sprachmodellen . . . 19
- 4.2 Datenkuration . . . 22
- 4.3 Unüberwachtes Weitertrainieren . . . 22
- 4.4 Ausführen des Training-Programme . . . 22
- 4.5 Klausurfragen . . . 22
- 4.6 Modellvergleich . . . 22

5 Ausführung der Lösung 23

6 Ergebnisse 25

7 Diskussion 27

Zusammenfassung 29

Literatur 31

Appendix

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
PaLM	Pathways Language Model
BPE	Byte Pair Encoding
BERT	Bidirectional Encoder Representations from Transformers
NLP	Natural Language Processing
ELMo	Embeddings from Language Models
SNIK	Semantisches Netz des integrierten Informationsmanagements im Krankenhaus
BeLL	Besondere Lernleistung

1 EINLEITUNG

1.1 GEGENSTAND

Eine effektive und effiziente Wissensbeschaffung bildet einen fundamentalen Bestandteil einer qualitativ hochwertigen klinischen Praxis in der Medizin. Bei jeder medizinischen Handlung werden große Mengen an Wissen genutzt und erzeugt, sei es als Grundlage für eine Diagnose oder zur Dokumentation des Behandlungsprozesses. Die strukturierte und klassifizierte Speicherung und Wiedergabe diesen Wissens 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 Semantisches Netz des integrierten Informationsmanagements im Krankenhaus (SNIK) (Jahn u. a., 2014), folgend 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

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 7.3.2023).

2 Institut für Medizinische Informatik, Statistik und Epidemiologie. https://www.imise.uni-leipzig.de/Institut (besucht am 9.3.2023).

medizinischen Anwendungen, eingesetzt, um neue Methoden der Wissensgewinnung 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 komplex, 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 Forschende, um diese Systeme zu erweitern oder neue Managementmethoden zu entwerfen. Eine konkrete und konsistente Wissensgrundlage 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 Arbeitsumgebungen. Da die verfügbare Literatur oft sehr umfangreich und ihre

Definitionen fragmentiert sind, stellt die Identifikation relevanten Wissens 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 Wissensbeschaffung, insbesondere für Studierende, die grundlegende Konzepte richtig verstehen wollen.

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

1.3 MOTIVATION

Eine Strukturierung von Wissen zum Management von Krankenhausinformationssystemen ist bereits Bestandteil des Projekts SNIK (Jahn u. a., 2014; Schaaf u. a., 2015), eine Ontologie des Instituts für Medizinische Informatik, Statistik und Epidemiologie der Universität Leipzig. Auf Basis dieser Ontologie wurden bereits verschiedene Methoden zur Wissensextraktion untersucht.
Die Besondere Lernleistung (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 Training 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 zusätzliches Training 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 soll 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 Verantwortungsbereich 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 Wissen über medizinisches Wissen, insbesondere über Informationssysteme im Gesundheitswesen, bereitstellt, unterliegt sie in ihrer Existenz als „Wissensbasis“ 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 Wissensbeschaffung 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 das extrahierte Wissen 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 Wissensextraktion 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 Training 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 wichtigen Wissens 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 Trainings 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

Wissen, Information, Daten (Winter) FScore, Precision, Recall Question Answering (Aggregation, erst FScore für Frage, dann Durchschnitt, oder andersrum) API

INHALTE DES KAPITELS

- Taxonomie von Ammus (Kalyan, Rajasekharan und Sangeetha (2022))
- Transformer Architektur (Vaswani u. a. (2017))
  - Neuronales Netz (https://katalog.ub.uni-leipzig.de/Record/0-1066754535)
- Activation Functions (https://katalog.ub.uni-leipzig.de/Record/0-1066754535) - Backpropagation (https://katalog.ub.uni-leipzig.de/Record/0-1066754535) - Training eines Modells - Batching - Optimizer - Learning Rate
  - Feed-Forward Netze - Multi-Head Attention - Encoder und Decoder
  - Input embeddings = Embeddings - Positional Embeddings - Tokenization - Byte Pair Encoding
  Sprachmodell Definition (Was ist das)
  - Autoregressive Modelle - Self-supervised Learning - Pretraining - continual Pretraining - Decoder-Based - ZeroShot / FewShot
  - Weiterentwicklungen - Deep Residual Connections - DropOut

2.1 TRANSFORMER

Die erste schriftliche Erwähnung des Transformer-Modells und zusätzlich auch die Einführung der beiden Teilmodelle Encoder und Decoder findet sich in Vaswani u. a. (2017). Die hier beschriebene bidirektionale Architektur bildet die Grundlage für alle darauf aufbauenden Modelle und Weiterentwicklungen. Die grundlegende Architektur wurde für verschiedene Anwendungen stark modifiziert. Seit 2017 gibt es grundlegende Unterschiede in den Modellen und deren Möglichkeiten. Aus diesem Grund haben Kalyan, Rajasekharan und Sangeetha (2022) eine Taxonomie der Transformer-basierten vortrainierten Sprachmodelle eingeführt. Diese Taxonomie wird hier zur Beschreibung weiterer Architekturen und Methodiken verwendet.

Neben dem Grundbaustein eines Transformers - dem Attention-Neuronales Netzwerk (NN) - sind zwei wichtige Modifikationen gegenüber normalen neuronalen Netzen in Transformer eingeflossen. Residuale Verbindungen als Level-Normalisierung, auf Englisch „De-

ep Residual Connections“, verändern das Ziel eines NN, behalten aber durch ihre Level-Normalisierung die gleichen Ausgaben bei. Dieses Konzept wurde erstmals in He u. a. (2016) eingeführt und liefert die Lösung für ein grundlegendes Problem von großen, aus mehreren Ebenen bestehenden Transformer-Modellen. Bereits 2016 wurde im Bereich der Bilderkennung festgestellt, dass sich die Korrektheit von Modellen mit zunehmender Tiefe sättigt und dann schnell verschlechtert, wenn dieses Modell weiter trainiert wird. Dies setzte eine praktische Grenze für die Tiefe von NNs und verhinderte somit die Lösung komplexerer Probleme mit größeren Modellen. He u. a. (2016) beschreiben eine Lösung durch die genannten Residuen, die normale NN simple ersetzen können, und zeigen ebenfalls die Wirksamkeit dieser Methode.

Die zweite wichtige Änderung ist die Einführung von Dropout. Dropout ist eine Methode, die 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. Diese Aussetzung hängt nicht von der Eingabe ab. Durch das Aussetzen von Neuronen wird das NN gezwungen, sich nicht auf andere Neuronen zu verlassen und somit eine bessere Generalisierung zu erreichen. Das Aussetzen erfolgt nur während des Trainings und nicht während der Inferenz. Die Methode wurde 2014 eingeführt und ist seitdem ein fester Bestandteil von NNs.

2.2 TOKENIZATION

Transformer-Modelle können Eingaben nicht ohne zusätzliche Umwandlung verarbeiten. Neben der Erzeugung von Kodierungsvektoren muss die Eingabe zunächst in kleinere Einheiten, sogenannte Tokens, zerlegt werden. Verschiedene ältere Modelle verwenden dazu Wörter oder Symbolunterteilungen. Dies ist jedoch problematisch.

Durch die Zerlegung der Eingaben in Symbole ist zwar das Vokabular kleiner, welches zu schnelleren Trainingsdurchläufen führt, jedoch muss das Modell vor dem Erlernen von Wortzusammenhängen, Satzstrukturen und Sachverhalten zunächst die Bedeutung der Wörter und deren Zusammensetzung aus Symbolen erlernen. Dies führt dazu, dass ein großer Teil der Trainingszeit für das Erlernen der Sprache verloren geht, was die endgültige Leistungsfähigkeit der Modelle massiv einschränkt (Sennrich, Haddow und Birch, 2016). Eine logische Schlussfolgerung wäre hier die Verwendung von Wörtern oder sogar Satzphrasen als Tokens. Mit zunehmender Größe der Datensätze, die zum Training der Modelle verwendet werden, wächst hier das Vokabular immens an. Dies führt zu einer starken Verlangsa-

mung der Trainingsläufe und zu sehr großen Modellen ohne Vorteil in ihrer Leistungsfähigkeit. Wörter mit gleichem Wortstamm oder ähnlicher Bedeutung aufgrund grammatikalischer Regeln (Plural, Genus, Tempora) müssen vom Modell erst als „gleiches Wort“ gelernt werden. Daher hat sich die Unterteilung von Wörtern in Teilwörter als Standard durchgesetzt.

2.2.1 *Byte-Pair-Encoding*

Sennrich, Haddow und Birch (2016) schlugen zu diesem Zweck die Verwendung von Byte Pair Encoding (BPE) vor. Die Unterteilung von Wörtern in Untergruppen von Wörtern hat bereits bei der Übersetzung von Sätzen zu erheblichen Verbesserungen geführt. Sie hat sich aber auch in anderen Bereichen und Aufgaben wie der Textgenerierung, der Textklassifikation und der Analyse von Emotionen durchgesetzt. Die Unterteilung von Wörtern ist hier eher als das Zusammenfügen kleinerer Teilwörter zu verstehen. Ausgehend von einem Vokabular, das aus allen Symbolen eines Alphabets besteht, wird dieses durch das Zusammenführen (engl. „Merge“) von Symbolen erweitert, deren Kombination im Datensatz am häufigsten vorkommt. Dieser Vorgang wird solange wiederholt, bis die gewünschte Anzahl von Teilwörtern erreicht ist. Die Anzahl der Teilwörter ist dabei ein Hyperparameter, der je nach Modell und Datensatz variiert.

Die Unterteilung von Wörtern in Teilwörter hat den Vorteil, dass die Größe des Vokabulars nicht mit der Größe des Datensatzes wächst. Dies führt zu einer schnelleren Eingabeverarbeitung und einer besseren Generalisierung der Modelle. Die Unterteilung von Wörtern in Teilwörter hat jedoch auch Nachteile. Sie ist nicht eindeutig, d.h. ein Wort kann in unterschiedliche Mengen von Teilwörtern zerlegt werden. Dies führt zu einer größeren Anzahl möglicher Eingaben, die das Modell lernen muss. Ein weiterer Nachteil ist, dass die Zerlegung von Wörtern in Teilwörter nicht immer sinnvoll ist. So kann es vorkommen, dass ein Wort in Teilwörter zerlegt wird, die in der Sprache nicht existieren. Dies wiederum minimiert die Verallgemeinerbarkeit der Modelle. Ein Beispiel hierfür ist das Wort „Datensatz“. Eine sinnvolle Unterteilung wäre hier „Daten“ und „satz“, aber durch den Aufbau des Vokabulars aus den Symbolen des Datensatzes kann es vorkommen, dass das Teilwort „Daten“ nicht die notwendige Häufigkeit besitzt und somit nicht im Vokabular vorhanden ist. Daher muss auch dieses Wort zerlegt werden, z.B. in „Da“ und „ten“. Beide Teilwörter haben in der deutschen Sprache keine Bedeutung, werden aber durch das Modell mit Bedeutung belegt und in Beziehung zu anderen Wörtern gesetzt. Dies führt zu einer unverständlichen Bedeutungsannotation von Teilwörtern und verschlechtert sowohl die

Leistung als auch die Nachvollziehbarkeit des Modells und erschwert die Forschung an den Modellen.

3 STAND DER FORSCHUNG

3.1 CONTINUAL PRETRAINING UND DIE NUTZUNG VON SPRACHMODELLEN

Ein Tranformer-Modell als Wissensbasis wird in Omar u. a. (2023) mit verschiedenen State of the Art (Stand der Technik) (SOTA)-Modellen verglichen. Sie zeigen eine deutliche Verbesserung der Robustheit gegenüber Eingabefehlern, der Erklärbarkeit von Antworten und des Fragenverständnisses bei komplexeren Fragen mit mehreren Fakten durch ChatGPT, zeigen aber Probleme bei der Aktualität von Wissen, dem Wissen über spezifische Domänen und vor allem bei der korrekten Beantwortung von Fragen. Der Grund dafür ist eine grundlegende Eigenschaft von GPT-Modellen, nämlich die fehlende Inkorporation aktuellen Wissens. Das Training von GPT-Modellen ist ein aufwendiger Prozess und kann nicht bei jeder Inferenz (Nutzung des Modells durch Generierung von Text) durchgeführt werden. Darüber hinaus wurde ChatGPT ohne zusätzliches Continual Pretraining eingesetzt. Eine Anpassung an die Domänen und eine Verbesserung der Korrektheit der Antworten stehen daher noch aus.

Radford und Narasimhan (2018) zeigen die Verbesserung der Leistung von Transformer-Modellen durch generatives Pretraining und eine weitere Verbesserung durch überwachtes Fine-Tuning. Auch hier ist ein klarer Trend erkennbar. Die Ergebnisse der Modelle verbessern sich mit zunehmender Größe des Datensatzes, mit zunehmender Länge des Trainingsprozesses und mit zunehmender Größe der Modelle.

Diese Tendenz wird durch Kaplan u. a. (2020) bestätigt, die hier den Einfluss verschiedener Einflussgrößen auf die Gesamtleistung eines Modells berechnen. Die hier verwendeten Einflussgrößen erlauben eine Vorhersage der Leistung eines Modells. Der Artikel schließt mit einer Abschätzung der theoretischen Maximalleistung und damit der Maximalgröße von Transformer-Modellen.

Um den beschriebenen Skalierungsregeln zu folgen, aber die Trainingszeit und die benötigte Datenmenge zu reduzieren, gibt es die Möglichkeit des Continual Pretraining. Gururangan u. a. (2020) hat diese Methodik angewandt und gezeigt, dass Modelle immens davon profitieren, domänenspezifisches Wissen zu adaptieren und aus der großen Menge an Basisdaten bessere korrekte Antworten in einer spezifischen Domäne zu generieren. Erstmals in Lee u. a. (2019) einge-

setzt, um das Basismodell Bidirectional Encoder Representations from Transformers (BERT) an die biomedizinische Domäne anzupassen, erweiterten Gururangan u. a. (2020) diese Methode und zeigten ihre Anwendbarkeit auf verschiedene Domänen und Aufgaben. Continual Pretraining auf aufgabenspezifischen Daten verbessert wiederum die Leistung für spezifische Aufgaben, während die Trainingszeit um den Faktor 60 kürzer ist im Vergleich zu Continual Pretraining auf Domänen. Eine Kombination beider Arten liefert die besten Ergebnisse.

Aber nicht nur das Continual Pretraining verbessert den Output der Modelle, sondern auch das überwachte Fine-Tuning. Ziegler u. a. (2019) beschreiben in ihrem Artikel die Effektivität von Reinforcement Learning als Fine-Tuning-Methode zur Lösung von Aufgaben der Weiterführung und Zusammenfassung von Texten. Fine-Tuning benötigt jedoch gekennzeichnete Daten (engl. „labeled data“, Daten mit bekannten korrekten Ausgaben), die in der Regel aufwendig zu erstellen sind und nicht immer in der benötigten Menge zur Verfügung stehen. In dieser Arbeit wird aufgrund der mangelnden Verfügbarkeit von gekennzeichneten Daten auf ein Fine-Tuning verzichtet.

3.2 AKTUELLE MODELLE UND DEREN NUTZBARKEIT

Die Erkenntnis, dass die Leistung von Modellen mit zunehmender Größe, Trainingszeit und Datenmenge steigt, hat zur Entwicklung einer Reihe von Modellen mit unterschiedlichen Architekturen, Anwendungsfällen und Leistungen geführt. Ein erster Durchbruch in der Leistungsfähigkeit von Transformer-Modellen wurde von Radford u. a. (2019) mit dem Modell GPT-2 erzielt. Im Vergleich zum ersten veröffentlichten GPT-1 Modell (Radford und Narasimhan, 2018) zeigten sie, dass Sprachmodelle Aufgaben lösen können, ohne explizit überwacht zu werden. Ebenso stellten sie fest, dass die Größe eines Modells, sei es hier die Anzahl der Parameter, die Größe des Datensatzes oder die Länge des Trainings, eine grundlegende Notwendigkeit für den erfolgreichen Einsatz der ZeroShot-Methode ist. Bereits hier konnten sie ohne jegliche Änderung der Architektur im ZeroShot Rahmen je nach Aufgabenstellung erfolgreiche, SOTA-kompetitive Ergebnisse erzielen.

OpenAI gelang mit GPT-3 ein Durchbruch in der Popularität und beschrieben ihren Ansatz in Brown u. a. (2020). In ihrem Artikel demonstrierten sie die Leistungssteigerung durch größere Modelle und zeigten, dass diese Leistung auch ohne Fine-Tuning erreicht werden kann. Sie verglichen auch das Antwortverhalten in Abhängigkeit von FewShot- und ZeroShot-Eingaben, wobei erstere bessere Ergebnisse lieferten. Diese Ergebnisse unterstützen die Hypothese, dass das in dieser Arbeit verwendete Modell auch ohne Fine-Tuning eine gute

Leistung erzielen kann.

Weiter im Jahr 2023 veröffentlichte OpenAI GPT-4 und stellten es in OpenAI (2023) vor. Neben deutlich besseren Ergebnissen durch ein noch größeres Modell mit mehr Parametern gelang es nun auch, Bilddaten als Eingabe zu verarbeiten. Dieser Artikel unterstreicht erneut die Annahme, dass größere Modelle eine bessere Leistung und ein besseres Verständnis der natürlichen Sprache haben. Eine Verwendung dieses Modells sowie des Modells GPT-3 ist nicht möglich, da diese Modelle derzeit nicht veröffentlicht sind.

Im Gegensatz zu den Modellen, die von OpenAI publiziert wurden, veröffentlichten Black u. a. (2022) GPT-NeoX. Ein Modell, das in Größe und Leistungsfähigkeit GPT-3 ähnelt, jedoch auf der Architektur von GPT-J[1] basiert und im Open Source Rahmen veröffentlicht wurde. Sie zeigten, dass die meisten interessanten Fähigkeiten eines Modells erst ab einer bestimmten Anzahl von Parametern sichtbar werden.

Zuletzt wurden von Touvron u. a. (2023) die Large Language Model Meta AI (LLaMa)-Modelle in verschiedenen Größen veröffentlicht. Ein klarer Vorteil gegenüber anderen Modellen in ihrer Anwendbarkeit ist hier der Fokus auf eine längere Trainingszeit und einen größeren Datensatz gegenüber der Modellgröße. Sie zeigten in fast allen Aufgabenbereichen bessere Ergebnisse als andere Modelle wie GPT-3 und Pathways Language Model (PaLM) mit deutlich weniger Parametern. Damit sind diese Modelle billiger, einfacher und schneller im Training und in der Anwendung mit gleichen oder besseren Ergebnissen. Diese Modelle wurden ebenfalls veröffentlicht und sind daher für diese Arbeit verfügbar.

3.3 FORSCHUNG UND PROBLEME VON MODELLEN

Neben der Entwicklung neuer Modelle wurden auch neue Ansätze zur Verbesserung des Continual Pretraining und der Adaption von Modellen entwickelt. Pfeiffer u. a. (2020) stellten in ihrem Artikel Adapter vor, die es ermöglichen, zusätzliche NNs auf verschiedenen Ebenen der Transformer-Architektur einzusetzen. Mit ihrer Hilfe kann die Adaption an andere Aufgaben und Domänen erreicht werden, ohne dass das gesamte Modell kontinuierlich vortrainiert werden muss, da während des Trainings alle Parameter des Ausgangsmodells fixiert bleiben, während die neu eingefügten Adapter trainiert werden. Darüber hinaus können bereits vortrainierte Adapter zu weiteren Domänen und Aufgaben in bestehende Modelle eingefügt werden, ohne dass ein erneutes Training erforderlich ist. Das veröffentlich-

1 Ben Wang https://github.com/kingoflolz/mesh-transformer-jax (abgerufen am 3.6.2023)

te System basiert auf dem Artikel von Houlsby u. a. (2019), in dem die Autoren das BERT-Modell auf 26 verschiedene Natural Language Processing (NLP)-Aufgaben trainierten, mit einer Anpassung von nur 3,6% der Parameter und einer Leistungsminimierung von 0,4% (ein ursprüngliches Training der Modelle auf diese Aufgaben hätte 100% aller Parameter angepasst). Sie bewiesen damit die Effizienz dieses Ansatzes, ohne die Leistung wesentlich zu beeinträchtigen.

Dai u. a. (2022) untersuchten die Fähigkeiten von Large Language Model (LLM)s im Hinblick auf ihre Fähigkeit, faktisches Wissen wiedergeben zu können, ohne eine Datenbank mit Fakten als Grundlage während des Betriebs zur Verfügung zu haben. Sie stellten fest, dass vor allem in tieferen Ebenen die neuronalen Netze so genannte „Wissensneuronen" besitzen, die mit bestimmten Fakten korrelieren. Diese Wissensneuronen werden aktiv, wenn ein bestimmter Fakt in der Eingabe angesprochen wird und können durch Verstärkung oder Unterdrückung dazu führen, dass das Modell diesen Fakt besser berücksichtigt oder „vergisst".

Die Untersuchung von Modellen und ihrer Fähigkeit, Sachverhalte zu erlernen und zu reproduzieren, wurde erstmals von Petroni u. a. (2019) vorgestellt. Die hier verwendeten Modelle BERT und Embeddings from Language Models (ELMo) wurden auf ihr Potential als unüberwachte offene Domäne QAS untersucht und zeigten gute Ergebnisse im Vergleich zu anderen SOTA-Systemen. Mehrsprachige Modelle wurden von Jiang u. a. (2020) auf die gleichen Eigenschaften hin untersucht, schnitten jedoch deutlich schlechter ab. Diese Ergebnisse deuten darauf hin, dass ein mehrsprachiges Modell für den Einsatz als QAS deutlich weniger geeignet ist, da ein Großteil der Leistung dieser Modelle in das Verstehen von Übersetzungen in andere Sprachen fließt.

Neben den überaus großen Erfolgen von neuen Modellen erheben sich jedoch auch neue Probleme bei der Anwendung dieser Modelle auf. Neben Falschaussagen ergeben sich Probleme durch soziale und andere Biases in den Antworten, Selbstüberschätzung bei Falschaussagen, die wiederum zu schwerwiegenden Problemen in der Anwendung dieser Modelle führen können, Generierung von schädlichen Inhalten, Unterstützung von Kriminalität durch Expertise und andere Probleme. Eine Analyse der Ergebnisse dieser Arbeit in Bezug auf die ethischen Richtlinien der GMDS findet sich in 1.5. OpenAI (2023) widmeten einen eigenen Abschnitt ihres Artikels der Untersuchung dieser Probleme und ihrer Adressierung. Sie zeigen darin grundsätzliche Probleme bei der Anwendung von Sprachmodellen auf, halten sich aber mit konkreten Lösungsvorschlägen zurück.

In Dehouche (2021) beschreibt der Autor weitere Fragen zum Umgang mit Antworten von Sprachmodellen und deren Konflikt mit dem Urheberrecht. Wem gehört der generierte Text - den Autoren der Datensätze, auf denen das Modell trainiert wurde, der Firma, der das Modell gehört, dem Benutzer, der das Modell anleitet? Auch hier zeigen sich ungelöste Probleme in der Anwendung von Sprachmodellen und bieten Raum für weitere Forschung. Eine abschließende Antwort auf diese Fragen gibt es noch nicht, so dass die Verwendung eines Sprachmodells zum Zeitpunkt dieser Arbeit nur von der Lizenzierung des jeweiligen Modells durch die Autoren des Modells und der Lizenzierung der Datensätze, die für das Continual Pretraining verwendet werden, abhängt. Das hier verwendete Buch Winter u. a. (2023) steht unter einer Open-Access-Lizenz und kann daher uneingeschränkt für ein Continual Pretraining verwendet werden.

4

LÖSUNGSANSATZ

VORGEHEN

- Rerchnernetz Aufbau und Konfiguration (warum reicht nicht eigener Computer)
- Auswahl Autoregressives Modell - Datenkuration Blaues Buch - Unüberwacht Weitertrainiert = Continual Pretraining - Nutzung von HuggingFace Trainer - Parallelisierung auf Rechennetz
- Auswahl Beispielklausuren - Beantwortung von Fragen
- Modellvergleich - Omar folgende Kriterien - Precision, Recall, F-Score als Benchmark - Modellvergleich GPT4 - Omar folgende Kriterien

4.1 AUSWAHL VON SPRACHMODELLEN

Um eine Auswahl eines Sprachmodells zu treffen, müssen verschiedene Kriterien erreicht werden. Basierend auf der Aufgabenstellung in 1.4 stehen unterschiedliche Modelle der Transformer-Architektur zur Auswahl. Desweiteren wird dies eingeschränkt auf Decoder-Basierende Modelle, da diese die Eigenschaften eines Autoregressiven Modells besitzen. Wie schon in **??** beschrieben, ist das Ziel eine Generierung von Text auf Basis einer Fragestellung. Diese Fragestellung steht am Anfang als Eingabe zur Verfügung, darauf folgend generiert das Modell kontinuierlich nachkommende Tokens. Neu generierte Tokens werden zusätzlich der ursprünglichen Fragestellung als Eingabe genutzt, um das darauf folgende Token zu generieren. Dieses Verhalten entspricht den Autoregressiven Modellen und benötigt Decoder-Basierte Modelle. Ein Decoder-basiertes Modell nutzt zu Generierung von Tokens (auch als Vorhersage von Tokens zu sehen) alle zuvor kommenden Tokens, während ein Encoder-Modell sowohl zuvor als auch nachstehende Tokens nutzt. Encoder-Modelle sind für die Text-Ausfüllung oder Text-Klassifikation gedacht, jedoch nicht für die Text-Generierung.

Desweiteren sind folgende Eigenschaften erwünscht. Das Modell sollte eine gute ZeroShot Leistung besitzen und nicht erst im FewShot oder OneShot Verfahren gute Leistungen erzielen. Diese Arbeit untersucht die Leistungsfähigkeit eines Modells, Wissen von domänspezifischer Literatur (hier von Winter u. a. (2023)) wiedergeben zu können. Durch das nutzen von zuvorigen Fragen und Antworten als Kontext, werden die Ergebnisse durch zusätzliches Wissen im Kontext synthetisch verbessert. Einige Fragen könnten nicht durch das Modell mit

Hilfe des erlernten Wissens beantwortet werden, sondern durch das Wissen im Kontext. Auch sollte eine gute Leistung des Modells ohne notwendiges Fine-Tuning vorhanden sein. Da ein Fine-Tuning in dieser Arbeit nicht möglich ist, begründet wegen der fehlenden Datenmenge, sollte ein Modell ausgewählt werden, welches ebenso gute Leistung ohne dieses Fine-Tuning erreicht.

Modelle die hier zur Auswahl stehend betrachtet werden sind GPT-2, GPT-3, GPT-4, GPT-J, GPT-NeoX und LLaMa. Modelle, die eine große Popularität gewonnen haben, jedoch nicht zur Auswahl stehen sind in Ausschnitten BERT, RoBERTa, DistillBERT, BART, T5, Opus, Pegasus, DialoGPT, Blenderbot und Flan-T5. Möchte man die Aufgabenstellung des Question Answering lösen, ist eine erste Wahl BERT-basierende Modell. Modelle wie BERT, RoBERTa und DistillBERT besitzen durch ihre Encoder-Architektur eine optimale Eigenschaft der Informationsextraktion aus Text. Jedoch ist dieser Ansatz limitiert in zwei Punkten. Die Modelle lernen hier nicht die Fakten und können neue Antworten auf Basis einer Fragestellung formulieren, sondern finden Textstellen eines Kontextes, welche die Fragestellung beantworten. Hier findet keine Textgenerierung statt, sondern es werden Auschnitte aus einem gegebenen Text-Kontext als Antwort gefunden. Dieser Kontext ist wiederum durch die Größe des Modells limitiert und beschränkt sich auf 512 Tokens. Ist die Antwort auf diese Fragestellung nicht im Kontext enthalten, kann das Modell keine Antwort finden. Da die Fragestellungen unabhängig vom Kontext beantwortet werden sollen, können diese Modelle nicht genutzt werden. Andere Modelle wie BART, T5, Opus, Pegasus, DialoGPT, Blenderbot und Flan-T5 sind für andere Aufgabenstellungen gedacht und können nicht auf die gegebene Aufgabenstellung angwendet werden. Sie sind für die Text-Klassifikation, Text-Übersetzung, Text-Zusammenfassung, Dialoge oder Text-2-Text Aufgaben gedacht.

GPT-2 ist das kleinste Modell der GPT-Reihe von OpenAI und wurde unter Radford u. a. (2019) erstmalig vorgestellt. Durch die geringe Größe kann das Modell auf einzelnen GPUs trainiert und genutzt (inferiert) werden, welches den Gebrauch des Modells deutlich vereinfacht. Auch ist es frei verfügbar und kann ohne weitere Maßnahmen genutzt werden. Jedoch ist die ZeroShot-Leistung nicht ausreichend. Das Modell erreicht eine Nachahmung der Texte, jedoch kein Verständis und Wiedergabe von Text.

GPT-3 und das darauf aufbauende GPT-3.5 (ChatGPT), vorgestellt in Brown u. a. (2020), lässt sich kostenlos zum Zeitpunkt der Arbeit nutzen, verfügt über eine sehr gute ZeroShot-Leistung und besitzt die Notwendige Größe um als Question-Answering Modell zu fungieren. Jedoch ist das Modell nicht frei verfügbar. Zwar existiert eine

api! (**api!**), jedoch dient diese nur zur Interferenz des Modells, ein Training ist nicht möglich. Damit fällst das Modell bereits aus der Auswahl. Ein weiteres Problem ist die Größe des Modells. Mit ungefähr 175 Milliarden Parametern ist eine Nutzung des Modells erst mit 8 A100 GPUs möglich. Das Trainieren benötigt zusätzlich ungefähr 4 mal so viel Leistung. Diese Leistung ist zwar durch das Rechenzentrum der Universität Leipzig zu bringen, verhindert allerdings die Nutzung des Modells in einer langfristigeren Umgebung.

GPT-4 ist das leistungsstärkste Modell von OpenAI und wurde in OpenAI (2023) vorgestelt. Auch dieses Modell ist nicht frei verfügbar und kann nur über eine **api!** genutzt werden. Die Leistung von GPT-4 ist jedoch gegenüber GPT-3 immens größer, weshalb die Untersuchung dieses Modells als Ersatz für ein eigens weitertrainiertes Modell angesehen ist. Die Größe des Modells wurde nicht von OpenAI veröffentlicht, liegt allerdings definitiv über GPT-3, weshalb hier eine Nutzung allein wegen der Größe ausgeschlossen ist.

GPT-J erreicht die Größe des kleinsten GPT-3 Modells mit 6.7 Milliarden Parametern, zeigt jedoch deutlich bessere Leistungen gegenüber GPT-2. Das Modell ist unter **gptj** veröffentlicht und frei verfügbar. Die Leistung des Modells ist im Vergleich zu anderen Modellen in der Auswahl geringer und wird deshalb ausgeschlossen.

GPT-NeoX wurde in Black u. a. (2022) vorgestellt und erreich mit einer Größe von 20 Milliarden Parametern die Leistungen von GPT-3 mit 175 Milliarden Parametern. Das Modell ist frei verfügbar und kann ohne weitere Maßnahmen genutzt werden. Die Leistung des Modells sind vergleichbar mit GPT-3, während die Größe des Modells ein Training mit 8 A100 GPUs ermöglicht. Diese Leistung wird jedoch von LLaMa übertroffen und wird deshalb nicht ausgewählt.

- Modelle mit speziellen Eigenschaften - Transformer-Architektur - Decoder-basiert - Autoregressiv - Gewichte sind verfügbar
- Modelle mit gewünschten Eigenschaften - gute Zero-Shot Performance - ohne Fine-Tuning nutzbar - überschreitet nicht Kapazität des Rechnernetzes
- Modelle zur Auswahl - GPT-2 - GPT-3 - GPT-4 - GPT-J - GPT-NeoX - LLaMa
- Modelle die nicht zur Auswahl stehen - BERT-basierend (Text Klassifikation): BERT, RoBERTa, DistillBERT - ZeroShot Classification: BART - Übersetzung: T5, Opus - Zusammenfassung: Pegasus - Dialogbasierend: DialoGPT, Blenderbot - Text-2-Text: FLan-T5
- Auswahlentscheidung - GPT-2: klein, einfach zu trainieren, öffentlich verfügbar, keine gute ZeroShot Performance - GPT-3 & GPT-4: groß, sehr gute Zeroshot Performance, nicht öffentlich verfügbar,

überschreitet Kapazität des Rechnernetzes - GPT-J: zwischen GPT-2 und GPT-3, öffentlich verfügbar, keine gute ZeroShot Performance - GPT-NeoX: groß, sehr gute Zeroshot Performance, öffentlich verfügbar, überschreitet Kapazität des Rechnernetzes - LLaMa: verschiedene Größen - optimale Größe auswählbar, sehr gute Zeroshot Performance, öffentlich verfügbar*, wird ausgewählt

4.2 DATENKURATION

- Formate die das Modell nicht verarbeiten kann - Bilder

Textpassagen ohne Wissen oder Kontext: - Bild-Beschreibun - Seitenzahlen - Vorwort - Autorenvorstellung

Optionale Textentfernung mit potentiell besserer Leistung: - Inhaltsverzeichnis - Literaturverzeichnis - Stichwortverzeichnis - Überschriften

4.3 UNÜBERWACHTES WEITERTRAINIEREN

- HuggingFace Trainer - Parallelisierung auf verschiedenen GPUS - Kommunikation zwischen GPUS mit Paket

- Alternativen zu Trainer - PyTorch - PyTorchLighning - Tensorflow

- Alternativen zu SLURM: keine, vorgegeben

4.4 AUSFÜHREN DES TRAINING-PROGRAMME

- Nutzung von SLURM - Kommunikation zwischen GPUS und Nodes - Long-Time Nodes - Speicherung von Modellen mit Checkpoints - Größe der Modelle = Anzahl notwendiger GPUs

4.5 KLAUSURFRAGEN

- Klausurfragen Antwortenerstellung - Umformulierung zu Prompt (da Finetuning nicht vorhanden)

4.6 MODELLVERGLEICH

- Berechnung von Precision, Recall, F-Score - Aggregation der Ergebnisse

Nutzung von Omar Kriterien: - Correctness - Determinism - Robustness - Explainability - Question understanding

Nicht genutzte Omar Kriterien - Incorporating recent Information - Generailty across different domains

- Vergleich der Modelle - Vergleich mit SOTA Modell - Nutzung von Kontext als Input?

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.

Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan und Sylvain Gelly (2019). „Parameter-Efficient Transfer Learning for NLP". In: *Proceedings of the 36th International Conference on Machine Learning*. Hrsg. von Kamalika Chaudhuri und Ruslan Salakhutdinov. Bd. 97. Proceedings of Machine Learning Research. PMLR, S. 2790–2799. URL: https://proceedings.mlr.press/v97/houlsby19a.html.

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.

Jiang, Zhengbao, Antonios Anastasopoulos, Jun Araki, Haibo Ding und Graham Neubig (Nov. 2020). „X-FACTR: Multilingual Factual Knowledge Retrieval from Pretrained Language Models". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, S. 5943–5959. DOI: 10.18653/v1/2020.emnlp-main.479.

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 mi-

ning“. 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].

Petroni, Fabio, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu und Alexander Miller (Nov. 2019). „Language Models as Knowledge Bases?“ 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)*. Hong Kong, China: Association for Computational Linguistics, S. 2463–2473. DOI: 10.18653/v1/D19-1250.

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.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei und Ilya Sutskever (2019). „Language Models are Unsupervised Multitask Learners“. In: URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

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.

Sennrich, Rico, Barry Haddow und Alexandra Birch (Aug. 2016). „Neural Machine Translation of Rare Words with Subword Units“. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, S. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://aclanthology.org/P16-1162.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever und Ruslan Salakhutdinov (2014). „Dropout: A Simple Way to Pre-

vent 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