Natürliche Sprachverarbeitung in Chatbots: Ein Literaturüberblick über aktuelle Ansätze und Transformer-Technologien

Bachelorstudium Informatik

Abgabe: [XX.XX.XXXX]

Inhaltsübersicht

1. Einleitung

Die rasant wachsende Präsenz von Chatbots in zahlreichen Bereichen, wie dem Kundendienst, der Bildung oder der Gesundheit, hat die Interaktion zwischen Mensch und Maschine verändert. Durch neue Methoden der künstlichen Intelligenz (KI) und aufgrund leistungsstärkerer Modelle haben die Chatbots in den letzten Jahren einen bemerkenswerten Fortschritt verzeichnen können und ihre Effizienz und Flexibilität entscheidend verbessert. Welche Technologien ermöglichen diese Entwicklungen? Und welche Herausforderungen sind damit verbunden? Die natürliche Sprachverarbeitung (NLP) spielt hierbei eine zentrale Rolle, da sie es Maschinen ermöglicht, Sprache zu verstehen und zu generieren. Transformer-Technologien, wie sie in BERT (Bidirectional Encoder Representations from Transformers) und GPT (Generative pre-trained transformer) genutzt werden, haben ein neues Kapitel in der Forschung und Nutzung der natürlichen Sprachverarbeitung aufgeschlagen.

In dieser Bachelorarbeit werden die modernen Transformer-Modelle in der natürlichen Sprachverarbeitung für Chatbots erforscht. Die Arbeit will beleuchten, inwiefern Transformer-Modelle einen neuen Stand in der Effizienz der Textverarbeitung und Kontextmodellierung geschaffen haben und was sie dazu befähigt hat, die Möglichkeiten rekurrenter neuronaler Netze (RNNs) oder LSTM (Long Short-Term Memory) zu überbieten. Das Ziel der Arbeit ist es zu ermitteln, inwiefern Transformer-Modelle die Entwicklung und Effektivität von Chatbots beeinflussen. Welche architektonischen Merkmale verleihen Transformer-Modellen ihre hohe Leistungsfähigkeit? Wie unterscheiden sich die Funktionalität und Architektur der Transformer-Modelle von den traditionellen NLP-Modellen? Und welche Vor- und Nachteile bieten diese beiden Ansätze in Bezug auf die natürliche Sprachverarbeitung für Chatbots? Zudem wird auch die hohe Komplexität der Transformer-Technologien und deren Auswirkungen auf den hohen Ressourcenverbrauch und die möglichen technischen Beschränkungen genauer betrachtet. Darüber hinaus wird auf wichtige Themen wie Datenqualität und ethische Bedenken der verwendeten Daten eingegangen. Welche Entwicklungen oder Trends in der Zukunft können erwartet werden und wie können sie die Fähigkeiten von Chatbots verbessern?

Die Arbeit basiert auf einer Literaturrecherche und kritischen Analyse von wissenschaftlichen Veröffentlichungen und Studien im Zusammenhang mit Transformer-Modellen und NLP für Chatbots. Zudem wird die Funktionsweise, Architektur und Anwendungen der modernen NLP-Technologien kritisch bewertet, verglichen und synthetisiert. Außerdem werden die

Historie, Entwicklungen und zentrale Innovationen der NLP-Technologien recherchiert.

Die Rechercheergebnisse zeigen, dass aktuelle NLP-Systeme durch Transformer-Modelle wie GPT-4 oder BERT neue Standards bezüglich ihrer Fähigkeiten erreicht haben. Diese Modelle sind zum Beispiel in der Lage, Kontextabhängigkeiten mithilfe des Self-Attention-Mechanismus sehr effizient zu modellieren. Demgegenüber stehen aber einige noch vorhandene Herausforderungen wie die hohen Rechenkosten und der hohe Ressourcenbedarf sowie die Möglichkeit, dass es zu Verzerrungen durch fehlerhafte Trainingsdaten oder bei der Verarbeitung von langen Kontexten kommen kann.

Die vorliegende Arbeit ist in sechs Kapitel unterteilt, wie im Folgenden kurz dargelegt wird:

In der Einleitung wird kurz auf die Thematik der Arbeit eingegangen. In Kapitel 2 werden die Transformer-Technologien dargestellt und auf deren Funktionsweise eingegangen. Das Kapitel 3 dient als Verbindung zur Anwendung von NLP in Chatbots. Es wird erläutert, wie NLP-Technologien in Chatbots genutzt werden und welche Vor- und Nachteile sie im Vergleich zu klassischen Ansätzen bieten. In Kapitel 4 werden verschiedene Herausforderungen und Limitationen im Kontext von NLP für Chatbots analysiert. Das fünfte Kapitel beschäftigt sich mit den Entwicklungen und Trends der NLP-Technologie, die die Effektivität von Chatbots verbessern werden. In der Konklusion (Kapitel 6) werden zusammenfassend die Erkenntnisse der Arbeit dargestellt.

2. Grundlagen der Transformer-Technologie

Die Grundlagen der Transformer-Technologie beleuchten die Entwicklungshistorie, Architektur sowie Funktionsweise und zeigen den Bezug zur Arbeit auf.

2.1 Entwicklung und Evolution von Transformer-Modellen

Die Entwicklung moderner Sprachmodelle hat die natürliche Sprachverarbeitung im letzten Jahrzehnt nachhaltig verändert. Begonnen hat der Wandel 2013 mit der Veröffentlichung von word2vec, einem vortrainierten Modell mit 6 Milliarden Wörtern, das mit Hilfe vektorbasierter Wortrepräsentationen eine grundlegende Weiterentwicklung in der natürlichen

Sprachverarbeitung darstellte (Rosinski, 2022, S. 32). Auch wenn dieses Modell eine deutliche Verbesserung gegenüber früheren Methoden zur Modellierung von Wortbedeutungen brachte, wies es deutliche Limitationen in der Beschreibung kontextabhängiger Wortbedeutungen auf. Aus diesem Grund entwickelte man ELMo, ein vortrainiertes, kontextuelles Modell, das mit Hilfe von 3-Layer-LSTM und einer Milliarde Wörtern in der Lage war, Wortbedeutungen dynamisch und abhängig vom Kontext zu repräsentieren (Rosinski, 2022, S. 32). Aufgrund von Problemen in der sequenziellen Verarbeitung sowie der mangelnden Skalierbarkeit war das Modell allerdings nicht im großen Maßstab einsetzbar (Krüger, 2021, S. 2).

Mit der Veröffentlichung von Transformer im Jahr 2017 erfolgte eine weitere signifikante Weiterentwicklung in der NLP-Technologie. Die Transformer-Architektur nutzte das Konzept der Self-Attention. Dieses Konzept ermöglicht es, Relationen zwischen allen Wörtern eines Eingabetextes in parallelen Prozessen zu modellieren (Rosinski, 2022, S. 32; Krüger, 2021, S. 2). Der Self-Attention-Mechanismus ersetzt das Konzept der sequenziellen Modellierung, welches für die mangelnde Parallelisierung und die Limitation von Langzeitabhängigkeiten verantwortlich war. Neben den kürzeren Trainingszeiten (Lakew et al., 2018, S. 3; Krüger, 2021, S. 2) ermöglicht das Konzept eine fast unbegrenzte Anzahl an Kontextwörtern bei der Analyse und Modellierung von Texten und ist somit besonders nützlich für Chatbots. Außerdem wurde das Konzept sehr schnell auf andere Anwendungsbereiche wie maschinelle Übersetzung und Textklassifikation angepasst und somit für verschiedene Anwendungsfälle eingesetzt (Jagtap & Salunke, 2025, S. 2; Lakew et al., 2018, S. 6–7).

Das Transformer-Modell ist eine Weiterentwicklung der neuronalen Netze und führt im Bereich der Natural Language Processing zu besseren Ergebnissen als klassische NLP-Modelle. Beim IMDB-Datensatz zum Beispiel lag die Genauigkeit der Sentimentanalyse der Transformer bei 90,3 % im Vergleich zu 87,1 % bei BiLSTM. Auf dem Nachrichtenklassifikationsdatensatz (AG News) schnitt die Transformer-Architektur mit 93,6 % besser ab als die BiLSTM mit 91,2 %. Außerdem erreichte die Transformer-Architektur im WMT’14 EN-DE-Datensatz der maschinellen Übersetzung einen BLEU-Score von 28,9 (BiLSTM: 25,8). Zusätzlich zur Ergebnisverbesserung wurden auch die Trainingszeiten der Modelle deutlich verringert (Jagtap & Salunke, 2025, S. 2). Transformer-Modelle, zum Beispiel GPT von OpenAI oder BERT von Google, werden nun standardmäßig für viele anspruchsvolle NLP-Anwendungen, wie Chatbots, Textgenerierung oder Übersetzung, genutzt (Rosinski, 2022, S. 32; Lehmhaus et al., 2023, S. 4).

Die nächste wesentliche Weiterentwicklung war die Skalierung der vortrainierten

Transformer-Modelle. Diese Skalierung führte zur Entwicklung von Modellen wie GPT-3 von OpenAI (170 Milliarden Parameter) oder GShard von Google (600 Milliarden Parameter) (Rosinski, 2022, S. 32). Das Hauptproblem beim Training der Transformer-Modelle ist die dafür benötigte Hardware, die für das Training komplexer Modelle immer leistungsfähiger werden muss. Außerdem steigt mit der Modellkomplexität auch der Energieverbrauch sowie die Notwendigkeit, Fairness und Inklusivität zu gewährleisten (Lehmhaus et al., 2023, S. 8).

Die Transformer-Architektur ermöglicht auch, external vortrainierte Embedding-Matrizen hinzuzufügen und direkt in eigene, domänenspezifische, kontextsensitive Embeddings zu integrieren. Dies ermöglicht eine weitere, domänenspezifische Modellierung und ist insbesondere für die Entwicklung von Chatbots ein nützliches Merkmal. Außerdem zeigen aktuelle Forschungen, dass hybride Modelle, die maschinelle Lernverfahren mit wissensbasierten Methoden kombinieren, das Potenzial dialogorientierter Systeme verbessern können (Hinkelmann et al., 2025, S. 13).

Dennoch sind viele Herausforderungen im Bereich der Transformer-Technologie noch immer nicht gelöst. Die Skalierbarkeit und Robustheit von Transformer-Modellen bei diversen Daten bleiben weiterhin wichtig, und die Verzerrungsgefahr bei vortrainierten Modellen sowie die Gefahr von Diskriminierung sind große Probleme. Die Hardware-Anforderungen und der ökologische Fußabdruck des ressourcenintensiven Trainings sind weitere Aspekte, die in der Entwicklung und Nutzung von Transformer-Modellen eine große Rolle spielen (Lehmhaus et al., 2023, S. 8; Hinkelmann et al., 2025, S. 13).

2.2 Architektur und Funktionsweise

Die Architektur der Transformer-Technologien ist grundlegend für deren Performance in der natürlichen Sprachverarbeitung. Ein wichtiges und innovatives Merkmal von Transformer-Modellen ist der Self-Attention-Mechanismus, welcher Beziehungen zwischen Wörtern einer Sequenz unabhängig von der Position erfasst. Dadurch wird eine Schwäche von rekurrenten neuronalen Netzen (RNNs) minimiert, und zwar der Verlust von langfristigen Kontexten. Gerade für Anwendungen wie Chatbots, welche meist lange Textmengen verarbeiten müssen, in welchen semantische Informationen teilweise sehr weit auseinanderliegen, ist diese Kontextsensitivität von Vorteil für die Qualität der Gesprächsführung. Modellintern, also vor der Antwortvorhersage, werden Tokens mit Hilfe von Attention-Gewichten priorisiert. Die Gewichtung kann der Grund für die Genauigkeit der

Antworten sein, welche sich laut empirischen Ergebnissen positiv auf die Kundenzufriedenheit auswirkt (Lakew et al., 2018, S. 3; Caelen & Blete, 2024, S. 7).

Ein weiteres Hauptmerkmal von Transformer-Modellen ist die Multi-Head-Attention. In der Multi-Head-Attention wird der initiale Embedding-Vektor auf mehrere Vektorräume aufgeteilt. Das ermöglicht parallele Repräsentationen, unterschiedliche Kontexte und ermöglicht so die Erkennung mehrdeutiger Situationen in der Anfrage des Benutzers. Die parallele Verarbeitung sorgt auch für die Effizienz, welche nötig ist, um die großen Datensätze in großskaligen Transformer-Modellen wie GPT-3 und GShard zu trainieren. Durch eine schnellere Verarbeitung und genauere Interpretation komplexer oder mehrdeutiger Benutzeranfragen wird die Performance und Nutzerfreundlichkeit in skalierbaren Chatbot-Technologien stark verbessert (Krüger, 2021, S. 15, S. 29; Rosinski, 2022, S. 32).

Die Feed-Forward-Schichten werden nach jedem Attention-Block hinzugefügt. Sie sind voll verbundene, neuronale Netze, welche nichtlineare Transformationen in der Verarbeitung der Daten durchführen. Der Sinn dieser Schichten ist vor allem in der Antwortgenerierung von Chatbots, da Beziehungen und Muster so besser extrahiert und übertragen werden können. Auf diese Weise können in verschiedenen Bereichen (z. B. Jura oder Medizin) differenziertere Antworten auf komplexe Anfragen generiert und so nicht eindeutige Schlussfolgerungen gezogen werden (Rosinski, 2022, S. 32; Caelen & Blete, 2024, S. 6).

Ein großer Vorteil von Transformer-Technologien ist die Effizienz beim Training und die simultane Verarbeitung von Kontext. N-Gramm- und RNN-Modelle können im Kontextvergleich nur einen kurzen oder lokalen Kontext verarbeiten. Transformer-Technologien beziehen den ganzen Kontext bei der Vorhersage mit ein. Dies hat bewiesen, dass die Leistungen in verschiedenen Aufgaben wie maschineller Übersetzung und Sentiment-Analyse verbessert werden. Gerade bei komplexeren Unterhaltungen ist der erweiterte Kontext wichtig, um Referenzen oder Gesprächsinhalte, die früher passiert sind, miteinbeziehen zu können. Es konnte anhand einer Customer-Support-Anwendung auch bewiesen werden, dass die Nutzer*innenzufriedenheit mit Transformer-Chatbots um 6,2 % höher ist. Dies bestätigt die Behauptung, dass Chatbots, die auf Transformer-Technologien basieren, widerspruchsfreie Antworten generieren können und somit als natürlicher erscheinen (Lakew et al., 2018, S. 6–7; Caelen & Blete, 2024, S. 6).

Embeddings werden bei der Konstruktion von Transformer-Modellen als Matrizen implementiert. Damit kann ein Hybrid-Ansatz verwendet werden, bei welchem vortrainierte

Vektoren durch selbst trainierte ersetzt werden. Dies ermöglicht, die Transformer-Technologie auf domänenspezifische Anwendungen zu spezialisieren, wie zum Beispiel medizinische oder juristische Chatbots. Diese domänenspezifischen Embeddings können kombiniert mit den durch den Attention-Mechanismus erstellten kontextsensitiven Wortbedeutungen verwendet werden. So kann der Kontext von Informationen auch in domänenspezifischen Bereichen korrekt erfasst werden (Krüger, 2021, S. 9; Rosinski, 2022, S. 32).

Die Funktionsweise der Transformer-Technologie kann auch als Kaninchen-Metapher aufgezeigt werden, wie Anthrop (2024, S. 1) dies tut: Das „weiße Kaninchen“ steht für das Sprachverständnis, das „schwarze Kaninchen“ für die Sprachgenerierung und der „Hohlraum“ zwischen diesen beiden für die eigentliche Sprachverarbeitung.

Als Fazit kann man folgendes festhalten: Die Architektur von Transformer-Modellen hat die Leistung in der natürlichen Sprachverarbeitung deutlich verbessert. Der Self-Attention-Mechanismus und die Multi-Head-Technologien unterstützen das Lernen des Kontextes und der Bedeutung von Tokens. Diese beiden Punkte, kombiniert mit den oben genannten Eigenschaften, tragen zu einem vielversprechenden und akkurateren Einsatz von Chatbot-Technologien im Alltag bei.

3. Natürliche Sprachverarbeitung in Chatbots

Die Verarbeitung natürlicher Sprache durch Chatbots wurde durch den Wandel von klassischen Methoden hin zu Transformer-Technologien revolutioniert. Die Arbeit geht auf die verschiedenen Ansätze ein, angefangen bei klassischen Ansätzen, deren Weiterentwicklung zu Transformer-Architekturen und die Integration dieser Technologien in dialogorientierte Systeme. Das Ziel ist, die modernen Möglichkeiten zur Verarbeitung natürlicher Sprache und deren Anwendung in Chatbots zu verdeutlichen.

3.1 Traditionelle NLP-Ansätze

Die klassischen NLP-Ansätze bedienten sich einfacher Algorithmen zur Klassifizierung wie k-Nearest Neighbors (KNN), Naive Bayes und Maximum Entropy. Auf identischen

Datensätzen lieferten diese Methoden oft nur minimal differente Ergebnisse, was darauf hindeutete, dass die gewählten Merkmalsdarstellungen das Kernproblem darstellten und damit die größte Qualitätssteigerung der gesamten Sprachverarbeitung erzielt werden kann (Leser, 2024, S. 19). Naive Bayes erfreute sich aufgrund des linearen Modells großer Beliebtheit, da dies eine hohe Lern- und Speichereffizienz zur Folge hatte (Leser, 2024, S. 32). Empirische Ergebnisse haben allerdings gezeigt, dass klassische Verfahren mit starken Einschränkungen zu kämpfen hatten, sofern Kontexterkennung oder die Modellierung komplexer semantischer Beziehungen erforderlich waren (Leser, 2024, S. 19, S. 32). Dies machte sie ungeeignet für dialogorientierte Chatbots, die auf die richtige Anwendung von Polysemie und Synonymie angewiesen waren.

Die limitierte Ausdrucksstärke von klassischen N-Gramm-Modellen zeigt sich vor allem bei längeren oder komplex verschachtelten Texten. Da der Algorithmus sich auf das lokale Kontextfenster beschränkt, wird der übrige Textzusammenhang ausgeklammert (Rosinski, 2022, S. 32). Dies resultiert in grammatikalisch korrekten, aber inhaltlich falschen oder ungeeigneten Antworten bei einem Chatbot. Um auf einen sinnvollen, thematisch konsistenten Dialog zurückzugreifen, ist die Kontext-basierte Modellierung unabdingbar. Hier setzten sich Transformer-Modelle durch, welche den gesamten Sequenzkontext durch den Self-Attention-Mechanismus beachten konnten und so mehr Flexibilität für die Dialogmodellierung aufweisen.

Auch im multilingualen Raum zeigte sich der Trend zur Kontext-basierten Modellierung und brachte einige Vorteile mit sich. Um mehrere Sprachen einbinden zu können, müssen sämtliche Informationen bezüglich des sprachlichen Kontextes erhalten bleiben. Hier versagten traditionelle rekurrente neuronale Netze (RNN) sowie LSTMs auf ganzer Ebene. Mit dem Self-Attention-Mechanismus der Transformer wurde es möglich, Rechenschritte zu parallelisieren und Zero-Shot-Inferenz für bisher ungesehene Sprachpaare anzubieten (Lakew et al., 2018, S. 1, S. 3). Die Einführung der Transformer ermöglichte es daher erstmals, Chatbots für den internationalen Raum auf ein effizientes und realistisches Level zu heben. Diese Entwicklung wurde auch auf empirischer Ebene belegt, welche zeigte, dass Transformer-Modelle in der Lage sind, das domänenübergreifende Sprachverständnis zu verbessern. Bei RNN und LSTMs wurden stets große Mengen an domänenspezifischen Trainingsdaten benötigt (Lakew et al., 2018, S. 7).

Die früher verbreiteten regelbasierten Systeme und Dialogbäume arbeiten deterministisch, sind gut nachvollziehbar und bieten für Büroanwendungen die geringste Komplexität (Meyer von Wolff & Schumann, 2018, S. 20). Allerdings erweist sich das Hinzufügen neuer Regeln

als zeitaufwendig und komplex. Zudem sind diese Systeme nicht lernfähig. Für Chatbots ist die Antwort also bereits vorab definiert und nicht kontextabhängig.

Ein weiterer Ansatz für NLP-Methoden bestand in statischen Vektorrepräsentationen von Wörtern. word2vec war in der Lage, semantische Ähnlichkeit abzubilden und bildete eine gute Grundlage für viele der darauf folgenden Ansätze (Rosinski, 2022, S. 32). Dieser statischen Darstellung fehlt jedoch der Kontext. Dadurch konnte der spezifische Sinn homonymer und mehrdeutiger Begriffe im Satzkontext nicht erkannt werden. Diese unkontextualisierten Wortembeddings können in Chatbots nur stark eingeschränkt angewendet werden. Modelle wie ELMo und BERT setzten nun auf Kontext-sensitives Lernen. Dadurch können ein und dasselbe Wort unterschiedliche Bedeutungen und Darstellungen annehmen.

Die Limitierung von traditionellen NLP-Methoden offenbarte sich auch bei anspruchsvolleren NLP-Aufgaben, die durch einen hohen semantischen und syntaktischen Modellierungsbedarf geprägt sind. Oft fehlte klassischen Modellen eine ausreichende Ausdrucksstärke, um nicht-triviale Muster und logische Beziehungen zu erkennen (Sanford et al., 2023, S. 1). Viele traditionelle NLP-Algorithmen haben Schwierigkeiten, auf algorithmischer Ebene Matching-Probleme oder Disjunktion zu bewältigen (Sanford et al., 2023, S. 1). Auch im Chatbot-Umfeld zeigte sich diese Fehleranfälligkeit in Form einer limitierenden Dialogtiefe oder in einer reduzierten User-Zufriedenheit.

3.2 Integration von Transformer-Technologien

Die Integration von Transformer-Technologien in Chatbots hat durch Self-Attention die Möglichkeit der natürlichen Sprachverarbeitung erweitert (Krüger, 2021, S. 2). Dieser ermöglicht es, Beziehungen zwischen allen Elementen der Eingabesequenz gleichzeitig zu betrachten, egal wie weit die Eingabepositionen voneinander entfernt sind. Da die Modellierung von Zusammenhängen in Dialogen besonders wichtig ist, kommt diese Technologie für Chatbots infrage (Krüger, 2021, S. 2). RNNs operieren sequenziell über die Eingabesequenz hinweg, während Self-Attention die gesamte Eingabesequenz gleichzeitig betrachtet und eine parallele Berechnung möglich macht. Das bringt Vorteile mit sich, wie z. B. geringere Trainingszeiten und kürzere Antwortzeiten (Lakew et al., 2018, S. 3). Dadurch sind Transformer-basierte Chatbots schneller, präziser und genauer als RNNs (Lakew et al., 2018, S. 3), insbesondere bei langen und komplexen Konversationen. Zudem lassen sich

die aktuellen Dialogbeiträge des Kunden mit vorangegangenen gewichten, was vor allem für kundenbezogene Dialoge von Nutzen sein kann, da sie dadurch stets kontextualisiert sind. Auch auf längere Sätze skalieren die Transformer-Modelle gut, wodurch der Chatbot einen Kunden in der gesamten Customer Journey unterstützen kann.

Moderne Transformer-Modelle sind im Gegensatz zu RNNs multilingual und bieten Vorteile beim Transferlernen. Da RNNs nur mit den Zieldaten für ein Sprachpaar trainiert werden, haben die Modelle eine geringe Flexibilität und können nur für ein einzelnes Sprachpaar verwendet werden. Im Gegensatz dazu nutzen Transformer-Modelle wie GPT und BERT einen gemeinsamen sprachübergreifenden Vektorraum und können ihr erworbenes Wissen von hochresourcigen auf low-resource Sprachen übertragen (Lakew et al., 2018, S. 1). Dadurch kann ein multilingualer Chatbot für diverse Sprachen und Sprachenpaare, ohne das Trainieren mit separaten Resourcen pro Sprachenpaar, eingesetzt werden. Durch Zero-Shot-Inferenz ist es möglich, einen Chatbot für neue Sprachpaare zu nutzen, die während des Trainings nicht gesehen wurden. Das macht sie im multilingualen Kundenservice einsetzbar und für Unternehmen besonders auf einem internationalen Markt interessant, da es eine kosteneffiziente und leicht skalierbare Möglichkeit bietet. Transformer-Modelle übertreffen in empirischen Benchmarks die Performance von klassischen RNN-Architekturen und bieten robustere Performance auf nicht-gesehenen Daten (Lakew et al., 2018, S. 1, S. 6–7).

Auch eine domänenspezifische Anpassung von Chatbots ist durch Transformer-Technologien besser möglich. Dabei sind die vortrainierten Wort-Embeddings ein wichtiger Punkt, um domänenspezifisches Wissen einzubinden (Krüger, 2021, S. 9). Die Verwendung von externen, vortrainierten Embeddings ermöglicht es dem Modell, die Bedeutungen besser zu verarbeiten, die für bestimmte Domänen wie Medizin, Steuerwesen oder das juristische Gebiet relevant sind. Dies führt zu verbesserten Antworten in diesen Domänen. Durch die Nutzung von zusätzlichen Embeddings kann der Chatbot auch mit klassischen Machine-Learning-Verfahren und klassischen Technologien wie OCR (Document Extractions) kombiniert werden, um die Vorteile beider Ansätze zu kombinieren und zu einem hybriden System zu gelangen. Dadurch kann man die Modellperformance bei einer geänderten Aufgabenstellung oder einem neuen Regelsatz optimieren, z. B. kann die Bedeutung des Tokens „Brutto“ im Sinne des steuerpflichtigen Bruttogehalts neu gelernt werden, oder neue Begriffe können schnell integriert werden (Beuther et al., 2024, S. 35). Ebenso kann man auch bei Änderungen des Vokabulars von synonymen Wörtern oder Abkürzungen wie z. B. Abkürzungen für Bundesländer schnell das vortrainierte Vokabular ändern und ergänzen. Außerdem werden branchenspezifische und unbekannte Akronyme

gelernt. Besonders für bestehende Systeme, in die Chatbots in bestehende Arbeitsabläufe integriert werden sollen, sind Kombinationen mit klassischen Technologien hilfreich. Zum Beispiel können bestehende OCR-Systeme zur Dokumentenextraktion eingebunden werden.

Ein weiterer großer Vorteil von Transformer-Modellen ist ihre Parallelisierbarkeit und damit Skalierbarkeit. Da sie eine parallele Verarbeitung der Eingabesequenz ermöglichen, verringert sich die Trainings- und Inferenzzeit (Lakew et al., 2018, S. 6). Durch diese verkürzte Trainingszeit lässt sich der Chatbot in vielen Anwendungsszenarien zügiger aufsetzen, was auch Unternehmen mit einem geringeren Budget zu Hause vor Ort in ihren Arbeitsalltag einsetzen können. Auch für sich ändernde Anforderungen oder für den Bedarf an einem Chatbot im Kundenservice, einer Steuerberatung für ihre Mandanten etc. kann diese Technologie einfach angepasst werden. Eine Studie zeigt, dass ca. 80–90 % der Standardfälle von dem Chatbot beantwortet werden können (Beuther et al., 2024, S. 31). Die kontinuierliche Aktualisierung der Modelle durch Feedbackschleifen im Betrieb sorgt für eine Steigerung der Leistung von Transformer-basierten Chatbots in der Praxis.

Herausforderungen bleiben jedoch, dass die Systeme für die jeweilige Situation angepasst werden müssen. Denn neben den technischen Anforderungen und dem Bedarf an verschiedenen Ressourcen muss sichergestellt werden, dass Chatbots eine gleichwertige Qualität in den verschiedensten Anwendungsfällen aufweisen können.

Die Data-Preprocessing-Schritte wie Tokenisierung und Schriftsystem sind wichtig. Untersuchungen haben gezeigt, dass Transformer-basierte Chatbots weniger gut für das Lesen von Devanagari ausgebildet sind als für das römische Skript (Neveditsin et al., 2024, S. 4). Die Anzahl der einzigartigen Tokens korreliert mit der Performance. Daher ist es wichtig, auf einen guten Tokenizer-Algorithmus zu setzen und diesen an die spezifischen Bedingungen (Sprache, Dialogformat etc.) anzupassen. Denn falsch verstandene Tokens, mehrdeutige Vokabeln und eine niedrigere Konsistenz können die Folge eines suboptimalen Tokenizer-Algorithmus sein.

Auch trotz des technologischen Fortschritts und der Möglichkeiten ist noch ein gewisser Weg zurückzulegen. Denn besonders im Bereich der Vertrauenswürdigkeit weisen viele User ein gewisses Maß an Skepsis auf (Baumann & Siegert, 2024, S. 30). Die große Heterogenität der Endnutzer*innen verlangt es, Chatbots inklusiv und nutzerzentriert zu gestalten. Um die heterogenen Bedürfnisse zu befriedigen, muss eine Vielzahl von Menschen verschiedener Altersgruppen, verschiedener Berufshintergründe und

unterschiedlicher technischer Versiertheit möglichst früh im Entwicklungsprozess miteinbezogen werden. In einer Studie zeigen viele befragte Expert*innen noch einen Aufholbedarf der Chatbots in Bezug auf Benutzerfreundlichkeit, Transparenz, Sicherheit und Zuverlässigkeit der Systeme (Baumann & Siegert, 2024, S. 30). Andere Studien weisen auf mangelndes Wissen über und mangelndes Vertrauen in Chatbots als Ursache für mangelhafte Akzeptanz hin. Gezielte Schulungen und transparente Informationen über Funktionsweise und Datenschutz von Chatbots können helfen, die bestehenden Bedenken und Vorbehalte zu verringern (Baumann & Siegert, 2024, S. 15).

Auch bei der Nutzung des Transformer-Modells gibt es eine Limitation. Zum Beispiel funktioniert die Kodierung von Algorithmusmustern in den Modellen nur bei einfacherer Kodierung (Sanford et al., 2023, S. 1). Die Ergebnisse einer empirischen Studie weisen darauf hin, dass Gradientenbasiertheit beim Training mit zu Problemen bezüglich der Reliabilität und Robustheit der Modelle führt (Sanford et al., 2023, S. 1). Dies kann in Anwendungsfällen mit sehr hohem Spezialisierungsgrad zum Problem werden, wie z. B. medizinische oder juristische Beratungen, in denen korrekte Ergebnisse geprüft werden müssen. Die Einführung eines alternativen Prüfungsmechanismus oder eines anderen Trainingsansatzes kann dafür genutzt werden, die Limitationen auszugleichen und die Robustheit zu garantieren. Jedoch kann man auch weiterhin versuchen, aktuelle Nachteile durch die kontinuierliche Erweiterung der Trainingsverfahren zu kompensieren und neue Evaluationstechniken entwickeln.

4. Herausforderungen und Limitationen

Trotz der beeindruckenden Fortschritte, die Transformer bereits erzielt haben, kommen auch viele technische Limitationen ans Tageslicht. Probleme wie Kontextfenster-Grenzen, hoher Ressourcenverbrauch oder algorithmische Grenzen beschneiden die Leistung und Skalierbarkeit von Transformer-Modellen. Dies ist ein wichtiger Aspekt der Reflexion auf den gesellschaftlichen und ökologischen Kontext.

4.1 Technische Einschränkungen

Die technischen Limitierungen von Transformer-Modellen sind trotz ihrer Innovationskraft

eine bedeutende Challenge für ihre Anwendbarkeit. Eine wichtige Beschränkung ist die fixe Fenstergröße, wie sie beispielsweise bei GPT-4 mit 32.000 Tokens gegeben ist. Damit können bei längeren Dialogen relevante Informationsabschnitte wegfallen und die Korrektheit der Antwort negativ beeinflussen (Sanford et al., 2023, S. 2; Seeser-Reich, 2025, S. 2). Das ist besonders bei Beratungs- oder Supportprozessen relevant, wo viele Interaktionen gespeichert werden, um darauf aufzubauen. Ansätze mit hierarchischen Speichersystemen und die Kontextunterteilung bieten jedoch eine gewisse Abhilfe (Seeser-Reich, 2025, S. 2). Hierbei entstehen jedoch wieder eine Reihe von Herausforderungen bei der algorithmischen Auswahl und Kontextoptimierung. Es gibt die Möglichkeit von hybriden Modellen, wo ein zusätzliches Speichermodul den externen Kontext bereitstellt, aber solche Ansätze sind derzeit eher experimentell.

Die hohen Ressourcenanforderungen von moderner Hardware sind eine Hürde für KMU, da spezielle GPUs oder TPUs und große Datenmengen in der Entwicklung solcher Modelle verwendet werden und eine spezielle Expertise für den Betrieb nötig ist (Seeser-Reich, 2025, S. 3; Cardoso et al., 2024, S. 3). Das kann sogar dazu führen, dass KMU in der Entwicklung und Anwendung von KI von größeren Unternehmen abhängig sind. Es ist notwendig, kostengünstigere Architekturvarianten zu schaffen oder KI-Dienste speziell für kleine Unternehmen anzubieten. Der Energieverbrauch ist auch ein Problem, das nicht nur in ökologischer Hinsicht zu sehen ist, sondern auch finanzielle Auswirkungen auf Betreiber und damit Anwender solcher KI-Systeme hat. Durch die Skalierung von Transformer-Modellen ist daher von einer zunehmenden Menge von CO_2-Emissionen auszugehen (Ecker & Dotzauer, 2023, S. 2). Dies sollte auch als Herausforderung in der Forschung betrachtet werden.

Es ist bekannt, dass Transformer-Modelle algorithmische Grenzen haben, zum Beispiel bei komplexen Matching- und disjunktiven Aufgaben. Die Grenze für solche Lernmodelle liegt auch im grundlegenden Lernmechanismus, der gradientenbasierte Algorithmen verwendet. Dies bringt Verzerrungen mit sich und erhöht die Anfälligkeit für Fehler (Sanford et al., 2023, S. 1; Chen et al., 2024, S. 1).

Ein großer Nachteil der meisten Modelle ist auch, dass sie extrem auf die Qualität der Trainingsdaten angewiesen sind. Fehlerhafte und unvollständige Daten führen zu sogenannten Halluzinationen und machen die Ergebnisse somit unzuverlässig. Damit sind Anwendungen im Kundenservice oder Gesundheitsbereich beispielsweise nur sehr eingeschränkt umsetzbar (Seeser-Reich, 2025, S. 3; Cardoso et al., 2024, S. 3).

Der Umgang mit kulturellen Unterschieden ist eine Herausforderung, die auf zwei Ebenen betrachtet werden kann. Einerseits können domänenübergreifende Transformer-Modelle durch Transferlernen in verschiedensten Anwendungsgebieten eingesetzt werden. Auf der anderen Seite können andere Schriftsysteme und viele verwendete Tokens ein Problem für die Performance darstellen (Lakew et al., 2018, S. 1; Cardoso et al., 2024, S. 3). Es ist eine Limitierung des aktuellen Stands der Technik, dass der Kontext vieler weniger ressourcenreicher Sprachen oder Dialekte nicht berücksichtigt wird, zum Beispiel, weil nicht genügend Trainingsdaten vorhanden sind und es nur wenige Tokenizer gibt. Dies könnte durch die Anpassung von Tokenizern und weitere Vorverarbeitungsschritte verbessert werden.

Die tieferen Transformer-Modelle sind oft rechen- und speicheraufwendiger, die breiteren Modelle haben oft Probleme bei der Komposition von mehreren Aufgaben (Chen et al., 2024, S. 4).

4.2 Ressourcenbedarf

Die Ressourcenanforderungen von Transformer-Modellen sind eine Herausforderung für die Praktikabilität der Technologie aufgrund der enormen Hardware, die erforderlich ist, um solche Modelle auszubilden und auszuführen. Um performante LLM-basierte Chatbots einsetzen zu können, ist ein RAM von mindestens 64 GB pro Server vonnöten, was Kleinunternehmen häufig nicht erreichen (Dauser & Utomo, 2025, S. 17). Der hierzu erforderliche Server stellt für viele KMU eine enorme Investition in die IT-Infrastruktur dar und somit eine Eintrittsbarriere.

Die wachsende Anzahl von Rechenzentren weltweit, wie zum Beispiel bei OpenAI, die Rechenzentren mit bis zu 5 GW planen, unterstreicht den Bedarf an leistungsfähigen KI-Lösungen (Brüggenwirth et al., 2024, S. 16). Dies zeigt einen wachsenden Fokus auf die Umweltauswirkungen und den Energieverbrauch von KI-Systemen.

Durch diese hohen Anforderungen bezüglich Entwicklung und Ausführung der Systeme werden kleinere Unternehmen in ihrer Fähigkeit, mit großen KI-Anbietern zu konkurrieren, eingeschränkt. In vielen Unternehmen sind große Technologieanbieter aufgrund der hohen Ausgaben im Einsatz, welche zum Teil monatlich bei 600 bis 2.000 € liegen (Dauser & Utomo, 2025, S. 17). Es handelt sich hierbei um reine Nutzungskosten, wobei die Kosten für

Updates, Lizenzgebühren, regelmäßige Wartung und Monitoring noch hinzukommen. Der Zugriff für kleinere Organisationen auf moderne Technologien muss in Zukunft erleichtert werden.

Offene Quellen stellen oftmals eine günstigere und freiere Lösung für moderne Chatbots dar, können aber im Vergleich zu marktdominanten Chatbots wie ChatGPT empirisch nicht mithalten (Dauser & Utomo, 2025, S. 11). Das macht einen Verzicht auf die eigene Unabhängigkeit notwendig und zementiert die Marktdominanz großer Unternehmen.

Um Transformer-Modelle zu trainieren, benötigt es enorme Datenmengen, wie das Beispiel von OpenGPT-X zeigt. Bei diesem Projekt wurden 2,5 Billionen Tokens aus dem Internet (80 %) und aus kuratierten Datensätzen (20 %) genutzt (Brüggenwirth et al., 2024, S. 15). Vorab werden in aufwendigen Filter- und Deduplizierungsprozessen große Teile der Daten entfernt, sodass am Ende noch ca. 60 % der Daten vorliegen. Dies erfordert erhebliche Rechenressourcen und stellt eine Herausforderung dar.

Die Qualität der zur Verfügung gestellten Trainingsdaten spielt hierbei eine essentielle Rolle und trägt maßgeblich zur Zuverlässigkeit und Genauigkeit der Modelle bei. Unpassende Datensätze können sogenannte Halluzinationen hervorrufen, also plausibel klingende und kontextgerechte, jedoch faktisch unzutreffende Antworten (Seeser-Reich, 2025, S. 3). Dies kann beispielsweise in Medizin- und Kundenserviceanwendungen ein Sicherheitsrisiko für Kundinnen und Kunden bedeuten. Bei Open-Source-Modellen spielen die Verfügbarkeit von Daten und die Möglichkeit zum Nachtraining eine wichtige Rolle.

Die Zuverlässigkeit kann erhöht werden, indem Nutzende die Möglichkeit haben, Feedback zu den Modellen zu geben (Dauser & Utomo, 2025, S. 10). Eine andere Herangehensweise kann auch sein, ein eigenes KI-Trainingsdatenset zu erstellen. So werden zum Beispiel im „Erfahrungsfeld-KI“ in Nürnberg Fortbildungen angeboten, um das notwendige Wissen zum Umgang mit KI-Systemen anzueignen (Bahlinger et al., 2023, S. 2).

KI-gestützte Chatbots bieten vor allem für KMU mehrere Vorteile, wie zum Beispiel die Kontrolle über Daten und die Anpassungsfähigkeit. Um dem Bedarf an zusätzlichen IT-Leuten zu begegnen, ist ein Umdenken nötig. So ist zum Beispiel das Outsourcing der IT möglich. Auch ist es ratsam, stets die aktuellen gesetzlichen, technischen und domänenspezifischen Anforderungen einzuhalten, da diese sich stetig ändern.

Die Anzahl der Parameter moderner Transformer-Modelle ist in den letzten Jahren rasant

gestiegen, so hat zum Beispiel GPT-1 117 Millionen, GPT-3 175 Milliarden und GPT-4 sogar ca. 100 Billionen (Brüggenwirth et al., 2024, S. 11). Das ermöglicht zwar sehr komplexe Systeme, führt aber auch zu einem steigenden Bedarf an ökologisch und ökonomisch nachhaltigen Konzepten. Da die Rechenzentren, die diese Systeme zum Laufen bringen, zu den größten Energieverbrauchern weltweit zählen, muss versucht werden, diese durch moderne, effiziente Hardware und ausgeklügelte Software zu verbessern. Auch ist es wichtig, sich mit der Frage zu befassen, wie sich der Energieverbrauch eines LLMs reduzieren lässt. Hier sind effizientere Trainingsansätze und Architekturverbesserungen der entscheidende Faktor. Im Marketing werden schon vermehrt AI-Assistants in Kombination mit klassischen Chatbots eingesetzt, wie zum Beispiel bei Klarna (Saum, 2024, S. 11). Hier muss aber noch ein intensiveres Umdenken stattfinden, um auch im Marketing eine zukunftstaugliche Technik einzusetzen (Seeser-Reich, 2025, S. 3).

Zusammenfassend ist die Nachhaltigkeitsdebatte nicht zu vernachlässigen und sollte in Zukunft ein Schwerpunkt bei der Konzeption und der Implementierung von Chatbot-Lösungen sein.

5. Zukünftige Entwicklungen

Der Blick nach vorn lässt erkennen, wie die Transformer-Entwicklung die Effizienz, Skalierbarkeit und Vielseitigkeit von Chatbots maßgeblich beeinflussen und so die Zukunft dieser Technologie aktiv vorantreiben kann. Jede Innovation und jedes neue Design birgt das Potenzial, die Limitierungen heutiger Modelle zu überwinden. Diese neuen Ansätze sind unabdingbar, um den steigenden Anforderungen an intelligente Sprachsysteme gerecht zu werden und eine nachhaltige Lösung in Bezug auf ökologische Aspekte zu schaffen. In diesem Kontext stellen die folgenden Ideen einen kleinen Vorgeschmack für die zukünftigen Möglichkeiten dialogorientierter KI-Systeme in der Arbeitswelt dar.

5.1 Optimierungspotenziale

Durch die Optimierungsmöglichkeiten für Transformer-Technologien lassen sich Verbesserungen für die Effizienz erzielen. Dabei wird das Problem in der Effizienzsteigerung des Self-Attention-Mechanismus identifiziert. Bei einer sehr langen Eingabesequenz werden

exponentiell viele Ressourcen im Rechenvorgang benötigt. Für die Analyse einer Wortsequenz aus nur zehn Wörtern sind 100.000 Einträge notwendig (Rosinski, 2022, S. 37). Lösungsansätze bieten eine adaptive Sparsity, wodurch nur die für den Self-Attention-Mechanismus notwendigen Werte aktiviert werden. Außerdem kann durch die Segmentierung der langen Wortsequenzen in kleinere Unterbrechungen ein Großteil der Datenverarbeitung gespart werden, ohne dass die Kohärenz verloren geht. Diese beiden Methoden bieten sich an, um die Anzahl der benötigten Ressourcen zu verringern. Hier muss das richtige Maß gefunden werden, damit die Qualität der Ergebnisse nicht darunter leidet.

Das zweite Optimierungspotenzial wird mit einer Erweiterung der nutzbaren Kontextfenstergröße angegeben, was die Leistung von Transformer-Modellen verbessert. Durch die Kombination von lokalen und globalen Attention-Mechanismen könnten umfangreichere Sequenzen bei längeren Dialogen betrachtet werden, ohne zu viele Hardwarekosten verursachen zu müssen (Rosinski, 2022, S. 37). Hier muss ein guter Kompromiss zwischen Komplexität der Implementierung und der zusätzlichen Hardware gefunden werden. Ein weiterer Ansatz ist die Verwendung von Speicherhierarchien, die oft verwendete Phrasen und Standardantworten speichern, was zur Reduktion des Speicherbedarfs beiträgt. In der Praxis wird dieses Modell zum Beispiel in Chatbots eingesetzt.

Als dritte Optimierung bietet sich ein hardwareseitiger Umstieg an. Mit Spezialchips wie TPUs oder GPUs können Vorteile in der Performance und Energieeffizienz von modernen Transformer-Modellen erzeugt werden, wodurch niedrigere Inferenzzeiten und weniger Betriebskosten die Folgen wären. Durch eine Optimierung des Resource-Managements auf Seiten der Software lassen sich ähnliche Vorteile herbeiführen. Eine potentielle Herausforderung der zukünftigen Forschung sind hier jedoch die ansteigenden Hardwarekosten, die durch verbesserte Software notwendig sein werden.

Der verbesserte Transformer durchgeführter Studien ermöglicht durch Veränderungen von Layer-Normalisierung, Residualverbindungen und Positionskodierungen enorme Leistungssteigerungen. Der entwickelte Enhanced Transformer führt gegenüber dem vorherigen Modell zu 202,96 % mehr Leistung auf BLEU-Score-Basis (Moon et al., 2023, S. 1, S. 6). Diese Methode bietet eine Perspektive für weitere Forschung auf dem Gebiet von Transformer-Architekturen. Die Erhöhung der Modellstabilität gelingt dem Enhanced Transformer durch Layer-Normalisierung, das für dialogorientierte Chatbots wichtig ist (Moon et al., 2023, S. 6). Zudem kommt es durch die Einführung von gewichteten

Residualverbindungen sowie durch reinforcement learning-basiertes Positions-Encoding zu Verbesserungen in der Textverarbeitung. Diese beiden Methoden ermöglichen es, zu entscheiden, welche Informationen verarbeitet werden müssen und welche Informationen ignoriert werden können.

Mit Hilfe von Zero-Masked-Self-Attention kann nicht-relevante Information aus dem Text maskiert werden. Dadurch werden die Ausgaben präziser und gleichzeitig wird der Informationsfluss verbessert (Moon et al., 2023, S. 6). Die Wirksamkeit in realistischen Anwendungen mit vielen Nutzern muss hier aber noch getestet werden.

Durch eine Kombination dieser Methoden könnten die neuen Generationen der Transformer-basierten Chatbots mit einer großen Bandbreite an verschiedenen Texttypen kommunizieren und sowohl korrekte als auch kreative Antworten liefern.

Die Verwendung von Transformer-Generatoren zusammen mit cWGANs ermöglicht das Ausgeben von realistischen Chatbot-Antworten. Hier werden Kohärenz und Konsistenz gesteigert (Esfandiari et al., 2023, S. 7). Im Experiment wird dabei ein vorab vortrainiertes Transformer-Modell als Basis für den Generator benutzt. Das Training durch cWGAN führte dabei zu besseren Antworten hinsichtlich Kohärenz und Relevanz (Esfandiari et al., 2023, S. 7). Beim Chit-Chat-Datensatz lieferte die Methode zudem einen höheren BLEU-4-Score (Esfandiari et al., 2023, S. 9). Ohne das adversariale Training fallen diese Werte im reinen Transformer deutlich niedriger aus, womit man schließen kann, dass es hier Synergieeffekte gibt. Eigene Ansätze könnten durch eine höhere Gewichtung von adversarials erzeugt werden oder aber durch die Verwendung domänenspezifischer Diskriminatoren, um Chatbots für den konkreten Bedarf weiterentwickeln zu können.

Die Optimierungspotenziale sind damit nicht nur durch strukturelle Änderungen im Model, sondern auch durch Optimierungspraktiken bei den Embeddings möglich. Die Verwendung einer einzelnen Embedding-Matrix für die Projektion der Embedding-Vektoren in kleinere Vektorräume mittels Gewichtsmatrizen ermöglicht die Verbesserung der Leistung moderner Transformer-Modelle (Krüger, 2021, S. 15, S. 29). Diese Verbesserungsmöglichkeit kommt in Anwendungsfällen wie Recht oder Medizin zum Tragen, in denen präzise Darstellungen notwendig sind. Eine weitere Möglichkeit ist die Verwendung eines Embedding-Moduls, in dem nicht nur Token aus einem Kontext, sondern die kombinierten Token aller Kontexte gespeichert werden. Diese hybride Vorgehensweise könnte neben dem semantischen Kontext der Chatbot-Eingaben auch emotionale Aspekte berücksichtigen. Damit lassen sich Modularisierungen erzeugen, die auf den Anwendungsfall abgestimmt werden können.

Transferlernen und Zero-Shot-Inferenz helfen bei der Multilingualität und Adaptierbarkeit von Chatbots. In ressourcenarmen Sprachen ermöglichen diese Methoden die Nutzung eines vortrainierten Systems von Hauptsprachen und das Wissen für die Sprachen zu übertragen, für die weniger Daten vorhanden sind (Lakew et al., 2018, S. 1, S. 3). Das Zero-Shot-Lernen erweitert darüber hinaus den Einsatzbereich von Chatbots außerhalb der im Training geübten Aufgaben und steigert die Skalierbarkeit und Effektivität. Dieser Aspekt wird zunehmend bei global agierenden Unternehmen und humanitären Organisationen eine wichtige Rolle spielen.

Durch eine Kombination von domänenspezifischem Transferlernen mit domänengenerischer Zero-Shot-Inferenz könnte die Konsistenz bei gleichzeitiger Förderung der Innovationsfähigkeit noch einmal erheblich gesteigert werden.

Insgesamt wird die Leistungsfähigkeit und Effektivität von Transformer-basierten Chatbots durch zahlreiche Optimierungen und Erweiterungsmöglichkeiten gesteigert.

6. Fazit

Die eingangs formulierte Zielsetzung der Arbeit, die modernen Transformer-Technologien für Chatbots im Bereich der natürlichen Sprachverarbeitung hinsichtlich ihrer Potenziale, Limitationen und Herausforderungen zu untersuchen, ist umfassend erreicht worden. Einhergehend mit einer Darstellung zentraler Innovationslinien seit der Einführung der Transformer und einer kritischen Reflexion von technischen und gesellschaftlichen Rahmenbedingungen konnte ein vielfältiges Bild gezeichnet werden. Angesichts der einleitend formulierten Forschungsfragen, welche wesentlichen Neuentwicklungen seit der Transformer-Modelle eingeführt wurden, welche wesentlichen Verbesserungen sich gegenüber klassischen Methoden erzielen lassen und welche Herausforderungen weiterhin in technischer als auch gesellschaftlicher Hinsicht zu lösen sind, konnte die Arbeit präzise Ergebnisse formulieren und in einen aktuellen wissenschaftlichen Kontext bringen.

Transformer und ihr auf dem Self-Attention-Mechanismus basierendes System haben die natürliche Sprachverarbeitung (NLP) wesentlich verbessert. Durch die Möglichkeit, ganze Textsequenzen parallel zu verarbeiten, kann auf dem System unabhängig von der Länge des zu bearbeitenden Textes über seinen Kontext nachgedacht werden. Es wird hierbei der

gesamte Kontext ohne Rücksicht auf seine Länge simultan bei der Textbearbeitung beachtet. Eine weitere Neuerung im Bereich der NLP durch Transformer besteht darin, dass mit den neuen Modellen Informationen aus unterschiedlichen Datenquellen kombiniert werden können. Insgesamt lassen sich anhand der Entwicklung der Transformer-Modelle feststellen, dass Systeme in der Folge genauer arbeiten und ihre Anpassungsfähigkeit verbessert wurde. Ein Vergleich von modernen Transformer-Modellen mit klassischen N-Gramm-Modellen, Modellen mit Long Short Term Memory (LSTM) und rekurrenten neuronalen Netzen (RNN) zeigt, dass sie diese in der Analyse von Stimmungen, in der automatischen Übersetzung oder beim Dialog in allen betrachteten Metriken übertreffen. In Hinsicht auf Multilingualität beweisen Transformer-Modelle durch Transferlernen und die Nutzung von externen Embeddings sowie durch Fine-Tuning die Fähigkeit, die sprachlichen Anforderungen verschiedener Sprachen zu bewältigen. Darüber hinaus werden durch die Einbindung von Kontext in moderne Modelle in Bezug auf einzelne Textabschnitte sowie auf domänenspezifische Texte Verbesserungen erreicht. In ihrer Leistungsfähigkeit können Transformer-Modelle im Vergleich zu klassischen Modellen durch neue Verbesserungen der Architektur überzeugen. Bei modernen Transformer-Modellen, wie den sparsity-basierten Transformer-Modellen, zeigt sich, dass sie noch genauer und ressourcenschonender arbeiten. Durch alle beschriebenen Veränderungen werden Chatbots mit modernen Transformer-Modellen in der Lage sein, auch komplexe, fachspezifische Dialoge im professionellen Sektor zu ermöglichen.

Weiterhin existieren auch Schwächen und Limitationen in der Arbeit der Transformer-Modelle. Der begrenzte Kontext von Modellen ist immer noch ein zentrales Thema der Transformer-Modellforschung. Die Hardware für die Ausführung von Transformer-Modellen ist relativ teuer und ressourcenintensiv, da die Modelle sehr umfangreich sind und über komplexe Strukturen verfügen. In Hinsicht auf Daten lassen sich ebenfalls verschiedene Schwächen und Limitationen im Rahmen der Arbeit der Transformer-Modelle nennen, da sie hohe Anforderungen an Datenqualität stellen, da ansonsten Halluzinationen (Fehlinformationen) auftreten können. Zudem ist es komplex, für die Datenethik robuste Nutzer*innenperspektiven und eine Akzeptanz in der Gesellschaft zu gewährleisten.

Die Ergebnisse werden an bestehenden nationalen und internationalen Studien mit den jeweils abgeleiteten Leistungen von Transformer-basierten Modellen gemessen. Eine differenzierte Einordnung mit verwandten Arbeiten wird zu folgenden Kernerkenntnissen abgeleitet: Multilingualität, Zero-Shot-Inferenz, hybride Architekturen, der Vergleich hinsichtlich Praxistauglichkeit im Kontext realer Ressourcenausstattung und Anforderungen

zur Akzeptanz in der Praxis sowie die Auseinandersetzung mit Limitationen und Nachhaltigkeitsherausforderungen. Diese Kernerkenntnisse wurden vor dem Hintergrund einer kritischen Reflektion der Limitationen der Transformer-basierten Chatbots betrachtet. Die vorliegende Arbeit betrachtet diesen Komplex als ein Spannungsfeld zwischen technischer Innovation und Verantwortung.

Aus der vorliegenden Arbeit lassen sich die folgenden Forschungsempfehlungen für zukünftige Arbeiten ableiten: Effizienzsteigerung von Transformer-Modellen, Weiterentwicklung von domänenspezifischen hybriden Modellen, Verringerung des Ressourcenbedarfs sowie der daraus resultierenden ökologischen Belastung, quantitative Nutzer*innenstudien zur Evaluation der Akzeptanz, Integration von domänenspezifischem Kontextwissen, Evaluierung von Modellen mit weiteren Evaluationsmetriken, Verbesserung des Bereiches der Ethik (insbesondere hinsichtlich Validierung von Evaluationsergebnissen und Transparenz), Ethisierung von Chatbots und Interpretierbarkeit und Nachvollziehbarkeit von Ergebnissen. Es wird davon ausgegangen, dass an einer Verbindung zwischen Computerwissenschaft, Sprachwissenschaft, Ethik und Sozialwissenschaften die Möglichkeit zur Forschung und Weiterentwicklung von NLP-Systemen für dialogorientierte Chatbots vorhanden ist.

Schlussendlich kann der folgende Teil der Arbeit zusammenfassend festgestellt werden. Durch diese Arbeit wurde die eigene Motivation für die Auseinandersetzung mit innovativen Fragen im Bereich der Technologien im Kontext sozialem Wandel gestärkt. Gleichzeitig wurde auch klar, dass die Zukunft der Arbeit in Bereichen wie der NLP oder des dialogorientierten Systems, in welchem Chatbots entwickelt und eingesetzt werden, sowohl eine exzellente wissenschaftliche Arbeit als auch ein hohes Maß an Verantwortung in dieser Entwicklung benötigt. In diesem Zusammenhang möchte die vorliegende Arbeit ihren Beitrag leisten und einen Überblick über die aktuellen Entwicklungen der auf der Transformer-Technologie basierenden Chatbotentwicklung geben. Die abgeleiteten Erkenntnisse der vorliegenden Arbeit werden anschließend zur Versachlichung der Debatte über die zukünftigen Möglichkeiten der Chatbotentwicklung beitragen.

Literaturverzeichnis

Anthrop, C. 3. (2024). Generative Künstliche Intelligenz und Vorab Trainierte Transformer. https://www.eulenhaupt.com/legal_translation_revision/english_german_dutch/text/Claude%203%20%C3%BCber%20Vorab%20Trainierte%20Transformer%20%28%27GPT%27%29.pdf

Bahlinger, T., Zimmermann, R., & Roth, A. (2023). Erfahrungsfeld KI. Technische Hochschule Nürnberg. https://opus4.kobv.de/opus4-ohm/files/1191/nbg_erfahrungsfeld_ohmdok_001.pdf

Baumann, T., & Siegert, I. (2024). Sprachassistenten. VDE ITG Informationstechnische Gesellschaft. https://opus4.kobv.de/opus4-oth-regensburg/files/7129/Baumann_Sprachassistenten_proceedings.pdf#page=18

Beuther, A., Rombach, A., Stephan, S., Fettke, P., Köppe-Karkutsch, J., & Dönnebrink, M. (2024). Künstliche Intelligenz im Steuerbereich. Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI). https://wts.de/wts.de/KI%20Studie/2024-04-11_KI-Folgestudie.pdf

Brüggenwirth, S., Burchard, A., Fingscheidt, T., Hoos, H., Illgner, K., Junklewitz, H., Kaup, A., von Knop, K., Köhler, J., Kutyniok, G., Martin, R., Kolossa, D., Möller, S., Schlüter, R., Thulke, D., Schmitt, V., Siegert, I., & Ziegler, V. (2024). Large language models are transformers in artificial intelligence, industry, education, and society. VDE Verband der Elektrotechnik Elektronik Informationstechnik e. V. https://www.vde.com/resource/blob/2359152/bc0c7b6d8464dc8e285618b35b11caa7/vde-position-paper-large-language-models-data.pdf

Caelen, O., & Blete, M.-A. (2024). Anwendungen mit GPT-4 und ChatGPT entwickeln (2. Aufl.). dpunkt.de. https://www.assets.dpunkt.de/leseproben/14157/Leseprobe.pdf

Cardoso, H. d. S., Kusser, N., & Kieselstein, J. (2024). Einsatz von Künstlicher Intelligenz bei der wissenschaftlichen Literaturrecherche: Ein Überblick [Dissertation, Universitätsbibliothek Augsburg]. Universitätsbibliothek Augsburg. https://opus.bibliothek.uni-augsburg.de/opus4/files/113159/113159.pdf

Chen, L., Peng, B., & Wu, H. (2024). Theoretical limitations of multi-layer Transformer. arXiv. https://arxiv.org/pdf/2412.02975

Dauser, D., & Utomo, M. (2025). KI-Chatbots Marke Eigenbau?! Forschungsinstitut Betriebliche Bildung (f-bb) gGmbH. https://www.f-bb.de/fileadmin/Projekte/ZZBY/250204_ZZSUE_f-bb-online_KI-Experimentierraum.pdf

Ecker, B., & Dotzauer, A. (2023). Generative KI in Bild- und Videosynthese [Dissertation, Hochschule Landshut]. https://www.haw-landshut.de/static/ITZ/Bilder/Veranstaltungen/LLF/Beitraege_LL/2024/LL_4_2024_GenKI_Bild-Video_Ecker_Dotzauer.pdf

Esfandiari, N., Kiani, K., & Rastgoo, R. (2023). A Conditional Generative Chatbot using Transformer Model. Semnan University. https://arxiv.org/pdf/2306.02074

Hinkelmann, K., Hoppe, T., & Humm, B. G. (2025). Hybride KI mit Machine Learning und Knowledge Graphs. Springer Vieweg. https://doi.org/10.1007/978-3-658-44781-6

Jagtap, C. S., & Salunke, S. Y. D. (2025). A comparative study of transformer vs RNN model. International Research Journal of Modernization in Engineering Technology and Science, 07(04), 6588–6590. https://www.doi.org/10.56726/IRJMETS73761

Krüger, R. (2021). Die Transformer-Architektur für Systeme zur neuronalen maschinellen Übersetzung – eine popularisierende Darstellung. trans-kom, 14(2), 278–324. https://www.trans-kom.eu/bd14nr02/trans-kom_14_02_05_Krueger_NMUe.20211202.pdf

Lakew, S. M., Cettolo, M., & Federico, M. (2018). A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation. Proceedings of the 27th International Conference on Computational Linguistics, 641–652. https://aclanthology.org/C18-1054.pdf

Lehmhaus, L., Kränzler, C., & Börner, M. (2023). Große Sprachmodelle. Bitkom. https://www.bitkom.org/sites/main/files/2023-06/BitkomLeitfadenGrosse-Sprachmodelle.pdf

Leser, U. (2024). Text Classification. Humboldt-Universität zu Berlin. https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/teaching/archive/ws_1819/vl_maschsprache/07_classification.pdf

Meyer von Wolff, R., & Schumann, M. (2018). Einsatz von Chatbots am digitalen Büroarbeitsplatz der Zukunft (Arbeitsbericht Nr. 1/2018). Georg-August-Universität

Göttingen. https://publications.goettingen-research-online.de/bitstream/2/120458/1/Arbeitsbericht_StandderForschung.pdf

Moon, W., Kim, T., Park, B., & Har, D. (2023). Enhanced Transformer Architecture for Natural Language Processing [Dissertation, Korea Advanced Institute of Science and Technology (KAIST)]. https://arxiv.org/pdf/2310.10930

Neveditsin, N., Salgaonkar, A., Lingras, P., & Mago, V. (2024). Classification of Buddhist verses: The efficacy and limitations of transformer-based models. Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, 377–385. https://aclanthology.org/2024.nlp4dh-1.37.pdf

Rosinski, W. (2022). Natural language processing. Hochschule Mittweida. https://www.staff.hs-mittweida.de/~rosinski/transformer/2022/presentation.pdf

Sanford, C., Hsu, D., & Telgarsky, M. (2023). Representational strengths and limitations of transformers [Doktorarbeit, Columbia University]. https://arxiv.org/pdf/2306.02896

Saum, J. (2024). Chancen und Herausforderungen von KI im Marketing [Dissertation, CommuniBIT e.K.]. CommuniBIT. https://www.unternehmerinnenforum-niederrhein.de/wp-content/uploads/2024/06/Saum_KI-im-Marketing_Skript.pdf

Seeser-Reich, K. (2025). Chatbots: Meilensteine der Entwicklung und Ausblicke in die Zukunft. DGUV Forum, 3/2025, 16-18. https://forum.dguv.de/issues/RZ_S016-018_1.04_Chatbots.pdf

Plagiatserklärung

Ich versichere, dass ich diese Arbeit selbständig angefertigt und keine anderen als die angegebenen Quellen benutzt habe.

Alle Stellen, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, habe ich in jedem einzelnen Fall unter genauer Angabe der Quelle (einschließlich des World Wide Web sowie anderer elektronischer Datensammlungen) deutlich als Entlehnung kenntlich gemacht. Dies gilt auch für angefügte Zeichnungen, bildliche Darstellungen, Skizzen und dergleichen.

Die vorliegende Arbeit wurde hinsichtlich Titel, Fragestellung, Aufbau und Inhalt, oder in umfangreichen Teilen und Auszügen daraus, noch nicht in einem Studiengang an dieser, oder einer anderen Hochschule, zur Anrechnung von Leistungspunkten vorgelegt.

Ich nehme zur Kenntnis, dass die nachgewiesene Unterlassung der Herkunftsangabe als versuchte Täuschung bzw. als Plagiat gewertet wird.

XXXX, den XX.XX.XXX