Kapitelübersicht

Schwerpunkte + Quellen

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

Bachelorstudium Informatik

Inhaltsübersicht

1. Einleitung

2. Grundlagen der Transformer-Technologie

2.1 Entwicklung und Evolution von Transformer-Modellen

Zusammenfassung:

Detaillierte Darstellung der historischen Entwicklung und kontinuierlichen Verbesserung von Transformer-Modellen, beginnend mit den ersten Implementierungen bis hin zu den neuesten Fortschritten und Versionen, die eine Grundlage für das Verständnis der aktuellen Technologien schaffen.

Schwerpunkte:

Detaillierte Darstellung der Entwicklungen von klassischen Sprachmodellen bis zu großskaligen Transformer-Modellen als Grundlage für moderne Chatbots, mit kritischer Einordnung der damit verbundenen Paradigmenwechsel, praktischen Fortschritte und anhaltenden Forschungsfragen

Die Entwicklung moderner Transformer-Technologien begann nach einer Phase, in der klassische Methoden wie word2vec von Google mit 6 Milliarden Trainingswörtern erstmals leistungsfähige, auf Vektoren basierende Wortrepräsentationen lieferten, aber durch die Einführung kontextueller Embeddings wie ELMo mit 3-Schicht-LSTM und schließlich durch den Paradigmenwechsel mit der Veröffentlichung des ersten Transformer-Modells 2017 abgelöst wurden, wobei Letzteres erstmals eine trainings- und skaliereffiziente Architektur auf der Grundlage des Self-Attention-Mechanismus etablierte (Rosinski, 2022, S. 32; Krüger, 2021, S. 2).

Die Self-Attention- und Multi-Head-Attention-Mechanismen der Transformer-Architektur ermöglichten es, im Gegensatz zu rekurrenten neuronalen Netzen (RNNs), Wortabhängigkeiten über große Distanzen und beliebige Positionen in der Sequenz effizient zu erfassen, wobei diese Architektur nicht nur die Genauigkeit in der Übersetzung und Textklassifikation deutlich steigerte, sondern auch die Trainingsgeschwindigkeit durch parallele Berechnung erheblich verbesserte (Lakew et al., 2018, S. 3; Jagtap & Salunke, 2025, S. 2; Krüger, 2021, S. 2).

Exemplarisch für den Fortschritt sind die Transformer-basierten Modelle GPT (OpenAI) und BERT (Google), wobei moderne Versionen wie GPT-3 mit 170 Milliarden Parametern und GShard mit 600 Milliarden Parametern die Skalierbarkeit und Leistungsfähigkeit auf ein zuvor unerreichtes Niveau gehoben haben – ein Trend, der sich in immer größeren Modellen und einer breiten Übertragung auf zahlreiche Anwendungsgebiete wie Chatbots, Textgenerierung und maschinelle Übersetzung widerspiegelt (Rosinski, 2022, S. 32; Lehmhaus et al., 2023, S. 4).

Die empirisch nachgewiesene Überlegenheit von Transformer-Modellen zeigt sich in praktischen Benchmarks und Anwendungen: Im IMDB-Datensatz erreichen sie z. B. eine Genauigkeit von 90,3 % gegenüber 87,1 % für BiLSTM, beim WMT'14 EN-DE-Übersetzungsdatensatz erzielte der Transformer einen BLEU-Score von 28,9 gegenüber 25,8 für BiLSTM und das Training verlief zudem schneller, was die

nachhaltige Ablösung der klassischen RNN-Ansätze durch Transformer-Architekturen in der Praxis begründet (Jagtap & Salunke, 2025, S. 2).

Transformer-Technologien haben mit ihrer Fähigkeit zur Generierung sprachlich kohärenter Texte und der Option, externe vortrainierte Word-Embeddings zu integrieren, den Grundstein für den Siegeszug großer Sprachmodelle und hybrider KI-Anwendungen gelegt, wobei aktuelle Forschung aufzeigt, dass die Kopplung von maschinellem Lernen und wissensbasierten Systemen neue Potenziale für die Weiterentwicklung eröffnet, insbesondere im Bereich komplexer, dialogorientierter Systeme wie Chatbots (Hinkelmann et al., 2025, S. 13; Krüger, 2021, S. 9).

Trotz dieser Fortschritte besteht weiterhin Forschungsbedarf im Hinblick auf die Skalierbarkeit, Robustheit und ethischen Herausforderungen von Transformer-Technologien – insbesondere im Kontext von Chatbots und generativen Sprachmodellen, da beispielsweise neben Hardware-Anforderungen auch Verzerrungen durch Trainingsdaten, limitierte Kohärenz bei der Textgenerierung und inklusive Sprache nach wie vor ungelöste Probleme darstellen (Lehmhaus et al., 2023, S. 8; Hinkelmann et al., 2025, S. 13).

Passende Quellen:

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

2.2 Architektur und Funktionsweise

Zusammenfassung:

Tiefgehende Erklärung der inneren Mechanismen und strukturellen Komponenten von Transformer-Architekturen, einschließlich Self-Attention, Multi-Head Attention und Feed-Forward Netzen, um ein umfassendes Verständnis der technischen Funktionsweise zu vermitteln.

Schwerpunkte:
Detaillierte Beschreibung des Self-Attention-Mechanismus als zentrales Element der Transformer-Architektur, das es ermöglicht, beliebige Abhängigkeiten zwischen Wörtern einer Sequenz unabhängig von deren Entfernung automatisch zu erkennen und zu gewichten, wodurch wesentliche Nachteile sequenzieller Modelle wie RNNs (z. B. Verlust von Langzeitkontexten bei langen Sätzen) überwunden werden (Krüger, 2021, S. 2; Lakew et al., 2018, S. 3; Caelen & Blete, 2024, S. 7).

Erläuterung der Multi-Head-Attention-Struktur, durch die die Informationsaufnahme in parallele, voneinander unabhängige Räume aufgeteilt wird, was sowohl die Modellierung komplexer Zusammenhänge als auch die Fähigkeit zur parallelen Berechnung und somit eine signifikante Beschleunigung des Trainings ermöglicht; typische Ausgestaltung: Aufteilung der ursprünglichen 512-dimensionalen Embedding-Vektoren in mehrere 64-dimensionale Vektorräume für die Query-, Key- und Value-Vektoren (Krüger, 2021, S. 15, S. 29).

Darstellung der Architektur von Feed-Forward-Schichten nach jedem Attention-Block im Transformer, die aus vollständig verbundenen neuronalen Netzen bestehen; diese sorgen für zusätzliche nichtlineare Transformationen und erhöhen die Ausdrucksstärke des Modells, was insbesondere für die erfolgreiche Generierung und Verarbeitung natürlichsprachlicher Antworten in Chatbots entscheidend ist (Rosinski, 2022, S. 32; Caelen & Blete, 2024, S. 6).

Kritische Einordnung der Trainingseffizienz und Kontextverarbeitung: Während klassische Modelle wie N-Gramm oder RNNs auf lokale Kontextfenster oder sequentielle Datenverarbeitung beschränkt waren, erlauben Transformer durch den Attention-Mechanismus die gleichzeitige Berücksichtigung des gesamten Sequenzkontextes, wodurch etwa in maschinellen Übersetzungssystemen oder bei Chatbots eine höhere Genauigkeit und Flexibilität erreichbar ist; empirisch zeigen Transformer konsistent bessere Ergebnisse als RNNs in Multilingual-NMT- und Benchmark-Tests (Lakew et al., 2018, S. 6–7; Caelen & Blete, 2024, S. 6).

Integration von Embedding-Matrizen, die für Transformer entweder speziell trainiert oder als externe, vortrainierte Matrizen genutzt werden können, wodurch die Architektur offen für hybride Ansätze aus klassischen und modernen Methoden bleibt; diese Flexibilität, kombiniert mit der Möglichkeit der Kontextualisierung von Bedeutung durch Attention, hat die Entwicklung maßgeschneiderter Lösungen für Chatbots und andere NLP-Anwendungen vorangetrieben (Krüger, 2021, S. 9; Rosinski, 2022, S. 32).

Anwendung der metaphorischen Modellbeschreibung nach Anthrop Claude 3: Die Transformer-Architektur umfasst die Rollen des codierenden „weißen Kaninchens“ (Verstehen), des dekodierenden „schwarzen Kaninchens“ (Generieren) und des dazwischenliegenden „Hohlraums“ als Sprachverarbeitung, was das Zusammenspiel der architektonischen Komponenten veranschaulicht und eine theoretische Grundlage für die Trennung und Kopplung von Sprachverständnis und Sprachproduktion in aktuellen Chatbots bildet (Anthrop, 2024, S. 1).

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

3. Natürliche Sprachverarbeitung in Chatbots

3.1 Traditionelle NLP-Ansätze

Zusammenfassung:

Überblick über die traditionellen Ansätze der natürlichen Sprachverarbeitung (NLP), ihre Methoden und Algorithmen, und deren Einschränkungen, die den Bedarf an fortschrittlicheren Techniken wie Transformer-Technologien verdeutlichen.

Schwerpunkte:

- Traditionelle NLP-Ansätze wie KNN, Naive Bayes und Maximum Entropy erzielen auf identischen Datensätzen oft nur geringe Leistungsunterschiede, wobei insbesondere Naive Bayes durch effizientes Lernen und geringen Speicherbedarf punktet – dies zeigt, dass die Wahl des Algorithmus, solange ähnliche Merkmalsrepräsentationen genutzt werden, oftmals keinen großen Einfluss auf die Ergebnisqualität hat (Leser, 2024, S. 19, S. 32).

- Die Effektivität klassischer Sprachmodelle, etwa N-Gramm-Modelle, wird durch die Beschränkung auf lokale Kontexte und die Unfähigkeit, langreichweitige Abhängigkeiten in Texten zu erfassen, fundamental limitiert; dies ist ein wesentlicher Grund, weshalb fortschrittlichere Methoden wie Transformer in der Praxis benötigt werden, um komplexe dialogbasierte Systeme wie Chatbots realistischer und flexibler zu gestalten (Rosinski, 2022, S. 32).

- In multilingualen oder ressourcenarmen Umgebungen zeigten klassische rekurrente neuronale Netze (RNNs) und ihre Varianten wie LSTMs Einschränkungen hinsichtlich Zero-Shot-Inferenz und Skalierbarkeit, da sie für neue Sprachkombinationen oder große Datenmengen wenig flexibel sind – insbesondere im direkten Vergleich mit dem später eingeführten Transformer-Ansatz, der diese Defizite überwindet (Lakew et al., 2018, S.

1).

- Die allermeisten klassischen Ansätze im Bereich der Chatbots basierten auf regelbasierten Systemen oder Dialogbäumen und dominierten lange Zeit vor dem Einsatz moderner KI-Technologien; diese Systeme waren zwar leicht nachvollziehbar, aber in ihrer Anpassungsfähigkeit und Skalierung für komplexe Aufgaben des digitalen Büroarbeitsplatzes extrem begrenzt (Meyer von Wolff & Schumann, 2018, S. 20).

- Frühere NLP-Modelle wie word2vec bildeten erstmals wettbewerbsfähige, vektorbasierte Wortrepräsentationen ab, konnten jedoch ohne Kontextualisierung in der Einbettung von Wortbedeutung keine zufriedenstellenden Ergebnisse für anspruchsvolle Anwendungsfälle wie dynamische Chatbots liefern – ein entscheidender Nachteil, der erst durch spätere kontextuelle Einbettungsmodelle und schließlich die Einführung der Transformer-Technologie überwunden wurde (Rosinski, 2022, S. 32).

- Untersuchungen zur Repräsentationsfähigkeit und Verzerrung klassischer Modelle, wie sie etwa durch einfache Algorithmen oder lineare Verfahren gegeben ist, belegen, dass traditionelle Ansätze algorithmischen Mustern oft nicht ausreichend folgen können und ihre Fähigkeit zur Modellierung komplexer semantischer Zusammenhänge im Vergleich zu späteren Transformer-basierten Architekturen empirisch limitiert bleibt (Sanford et al., 2023, S. 1).

Passende Quellen:

- 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
- 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
- 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

3.2 Integration von Transformer-Technologien

Zusammenfassung:

Ausführliche Beschreibung der Integration von Transformer-Technologien in moderne NLP-Systeme, wie sie traditionelle Methoden verbessern und die Leistung von Anwendungen wie Chatbots und Sprachassistenten steigern.

Schwerpunkte:

Effiziente Kontextualisierung und parallele Verarbeitung durch Self-Attention: Die Integration von Transformer-Technologien in Chatbots ermöglicht durch den Self-Attention-Mechanismus eine gleichzeitige Berücksichtigung aller Wörter einer Eingabesequenz, wodurch Abhängigkeiten zwischen weit entfernten Textelementen erkannt und genutzt werden können. Im Gegensatz zu klassischen RNN-basierten Systemen, die aufgrund ihrer sequentiellen Verarbeitung bei langen Kontexten an ihre Grenzen stoßen, steigern Transformer-basierte Chatbots sowohl die Antwortgenauigkeit als auch die Effizienz signifikant (Krüger, 2021, S. 2; Lakew et al., 2018, S. 3).

Leistungssteigerung und Multilingualität durch Transferlernen: Transformer-Modelle wie GPT und BERT ermöglichen es Chatbots, auf anspruchsvolle multilinguale Aufgaben zuzugreifen und sogar Zero-Shot-Inferenz durchzuführen – also Erkenntnisse auf Sprachpaare zu übertragen, die im Training nicht explizit vorhanden waren. Dies ist insbesondere für Chatbot-Anwendungen relevant, deren Einsatzgebiete verschiedene Sprachen umfassen und in ressourcenarmen Umgebungen robuste Lösungen erfordern (Lakew et al., 2018, S. 1).

Anpassungsfähigkeit an verschiedene Domänen durch flexible Embedding-Matrizen: Moderne Transformer-Architekturen erlauben die Nutzung und Kombination von speziell trainierten oder extern vortrainierten Embedding-Matrizen, wodurch Chatbots domänenspezifische Bedeutungen besser erfassen und verarbeiten können. Dies unterstützt die Entwicklung hybrider Systeme, die sowohl klassische als auch neuere Ansätze kombinieren, um etwa im Steuerkontext oder bei regulatorisch geprägten Aufgaben effektive Sprachverarbeitung zu gewährleisten (Krüger, 2021, S. 9; Beuther et al., 2024, S. 35).

Beschleunigte Lernprozesse und Skalierbarkeit für reale Anwendungen: Transformer-basierte Chatbots profitieren von der Parallelisierbarkeit des Trainings und der Modellstruktur, was nicht nur hohe Genauigkeiten, sondern auch eine schnellere Entwicklung und optimierte Anpassung an vielfältige Anwendungsszenarien wie Kundenservice oder Beratung ermöglicht. Die Einbindung von KI-Modellen in bestehende Arbeitsprozesse wird dadurch effektiver, was durch die Integration in Unternehmen sowie verschiedene Anwendungsstudien belegt wird (Beuther et al., 2024, S. 31; Lakew et al., 2018, S. 6).

Qualitative und quantitative Herausforderungen bei der Tokenisierung und Skriptverarbeitung: Experimentelle Untersuchungen zeigen, dass Transformer-basierte Chatbots bei der Verarbeitung bestimmter Schriftsysteme, wie etwa Devanagari gegenüber römischem Skript, an Einbußen in der Klassifikationsgenauigkeit leiden können. Die Anzahl der einzigartigen Token im genutzten Tokenizer korreliert dabei direkt mit der Modellperformance, was die Bedeutung einer optimalen Vorverarbeitung und Architekturwahl für spezifische Chatbot-Settings unterstreicht (Neveditsin et al., 2024, S. 4).

Praktische Akzeptanz von Transformer-Chatbots und Einfluss auf Nutzerverhalten: Untersuchungen zu Sprachassistenten und Chatbots verdeutlichen, dass trotz technologischer Fortschritte weiterhin Vorbehalte bezüglich Vertrauen, Sicherheit und Benutzerfreundlichkeit bestehen, was sich unmittelbar auf die Akzeptanz und Effektivität im Alltag auswirkt. Die Bandbreite im Interaktionsverhalten betont die Notwendigkeit einer integrativen Entwicklung, in die unterschiedliche Nutzergruppen einbezogen

werden, um Chatbots inklusiver und anwendungsorientierter zu machen (Baumann & Siegert, 2024, S. 15, S. 30).

Repräsentationsstärke und Limitationen im praktischen Einsatz: Obwohl Transformer-Modelle in vielen algorithmischen Grundoperationen Vorteile gegenüber klassischen Verfahren aufweisen, zeigen empirische Analysen, dass sie bei bestimmten Aufgaben, etwa der exakten Kodierung komplexer algorithmischer Muster, weiterhin inhärente Verzerrungen durch das Lernen aufweisen können. Diese Limitationen sind für Chatbot-Entwicklungen besonders relevant und müssen bei der systematischen Implementierung beachtet und durch gezielte Trainings- und Architekturanpassungen adressiert werden (Sanford et al., 2023, S. 1).

Passende Quellen:

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

4. Herausforderungen und Limitationen

4.1 Technische Einschränkungen

Zusammenfassung:

Analyse der technischen Limitationen von Transformer-Modellen, einschließlich Probleme wie Datenverzerrungen, Schwierigkeiten bei der Handhabung langer Sequenzen und die Herausforderungen der Modellskalierbarkeit.

Schwerpunkte:

Technische Limitationen bei der Handhabung langer Sequenzen und Kontextgrenzen: Transformer-Modelle wie GPT-4 sind in ihrer Fähigkeit, längere Kontexte zu berücksichtigen, durch eine fest definierte Kontextfenstergröße (z. B. bis maximal 32.000 Token) eingeschränkt, wodurch wichtige Kontextinformationen bei umfangreichen Dialogverläufen verloren gehen können und die Qualität der Chatbot-Interaktionen abnimmt (Sanford et al., 2023, S. 2; Seeser-Reich, 2025, S. 2).

Ressourcenintensive Trainings- und Betriebsanforderungen sowie deren Auswirkungen auf die praktische Verfügbarkeit: Die hohe Rechenleistung, der immense Speicherbedarf und der Bedarf an umfangreichen, hochwertigen Trainingsdaten stellen beim Einsatz von Transformer-Modellen wie GPT oder BERT eine zentrale Herausforderung dar, was nicht nur die Entwicklungskosten erhöht, sondern auch die Nutzung in kleineren Unternehmen oder ressourcenarmen Umfeldern erschwert; diese technische Barriere limitiert zudem die Aktualisierung und Weiterentwicklung von Chatbot-Systemen (Seeser-Reich, 2025, S. 3; Cardoso et al., 2024, S. 3; Ecker & Dotzauer, 2023, S. 2).

Empirische Verzerrungen und begrenzte algorithmische Repräsentationsfähigkeit: Transformer-Modelle zeigen bei bestimmten algorithmischen Grundaufgaben, wie komplexen Matching- oder Disjunktions-Problemen, Verzerrungen in der Repräsentation und Effizienz, insbesondere wenn diese mit gradientenbasierten Lernverfahren trainiert werden; diese Einschränkung kann zu fehlerhaften oder suboptimalen Ergebnissen in Chatbot-Anwendungen führen, wenn präzise semantische Zuordnungen oder logische Schlussfolgen erforderlich sind (Sanford et al., 2023, S. 1; Chen et al., 2024, S. 1).

Qualitätsprobleme aufgrund sensibler Abhängigkeit von Datenqualität und Kontextverständnis: Transformer-basierte Modelle neigen dazu, bei schlechter Datenqualität fehlerhafte oder ungenaue Ergebnisse zu liefern und können dabei sogar „Halluzinationen“ erzeugen – also überzeugend klingende, aber faktisch unzutreffende Aussagen –, was insbesondere im sensiblen Umfeld automatisierter Kommunikation erhebliche Risiken für Fehlinformationen und das Vertrauen der Nutzenden birgt (Seeser-Reich, 2025, S. 3; Cardoso et al., 2024, S. 3).

Einschränkungen beim Transfer auf spezifische Domänen sowie Herausforderungen bei Multilingualität und Tokenisierung: Auch wenn Transformer-Modelle durch Transferlernen grundsätzlich domänenübergreifend eingesetzt werden können, zeigen Experimente, dass Unterschiede im Schriftsystem oder eine hohe Anzahl an einzigartigen Tokens die Klassifikationsgenauigkeit beeinträchtigen können; dies erfordert eine gezielte Anpassung von Tokenizern und Datenrepräsentationen für jedes Chatbot-Szenario (Lakew et al., 2018, S. 1; Cardoso et al., 2024, S. 3).

Systematische Skalierungsprobleme und architektonische Grenzen bei Multi-Layer-Transformern: Theoretische Analysen belegen, dass die Effizienz und Repräsentationskraft von Multi-Layer-Transformern durch einen inhärenten Trade-off zwischen Modelltiefe und -breite limitiert werden, sodass für einige Aufgaben exponentiell viele Parameter benötigt werden und bereits kleine Veränderungen in der Modellarchitektur zu erheblichen Unterschieden in der Lösbarkeit führen können – dies legt nahe, dass weitere Architekturinnovationen für hochspezialisierte Chatbot-Anwendungen erforderlich sind (Chen et al., 2024, S. 1).

Passende Quellen:

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

4.2 Ressourcenbedarf

Zusammenfassung:

Erläuterung des hohen Ressourcenbedarfs für das Training und den Betrieb von Transformer-Modellen, einschließlich der Anforderungen an Rechenleistung, Speicher und Datensätze, um die praktischen Grenzen und Herausforderungen zu verdeutlichen.

Schwerpunkte:

- **Enormer Bedarf an Rechenleistung und Speicherressourcen als zentrale Hürde bei Training und Betrieb moderner Transformer-Modelle:** Für leistungsfähige Chatbots auf Basis von Large Language Models (LLMs) wie GPT-4 sind spezialisierte Hardware-Infrastrukturen mit mindestens 64 GB RAM pro Server erforderlich, was insbesondere bei Anwendungen in kleinen und mittleren Unternehmen (KMU) schnell die wirtschaftlichen und technischen Möglichkeiten übersteigt (Dauser & Utomo, 2025, S. 17). OpenAI plant zudem Rechenzentren mit bis zu 5 GW Stromverbrauch, was die enorme Energiedichte und den Skalierungsbedarf unterstreicht (Brüggenwirth et al., 2024, S. 16).

- **Hoher finanzieller Aufwand für Entwicklung und Betrieb LLM-basierter Chatbots mit monatlichen Kosten zwischen 600 € und 2.000 € limitiert den Zugang für viele Organisationen und verstärkt Abhängigkeiten von marktführenden Anbietern:** Für Open-Source-Lösungen mit vergleichbarer Leistung wie ChatGPT sind massive Investitionen in Infrastruktur und laufende Aktualisierungen unerlässlich, während die entwickelten eigenen Modelle oft nicht mit den marktdominierenden Systemen mithalten können (Dauser & Utomo, 2025, S. 11, S. 17).

- **Immense Datenmengen für Trainingsprozesse führen zu Herausforderungen in Datenmanagement, Qualitätssicherung und ethischer Kontrolle:** Beispielsweise

werden im OpenGPT-X-Projekt für das Training mehr als 2,5 Billionen Token verarbeitet, wobei 80% aus Webdaten und 20% aus kuratierten Quellen stammen. Die Filterung und Deduplizierung dieser Daten reduziert das Datenvolumen zwar um insgesamt über 40%, ein erheblicher Teil der Rechenleistung wird jedoch allein für diese Vorarbeiten benötigt (Brüggenwirth et al., 2024, S. 15). Gleichzeitig stellt die Sicherung einer hohen Datenqualität eine Grundvoraussetzung dar, um sogenannte Halluzinationen und Fehlinformationen im Chatbot-Betrieb zu vermeiden (Seeser-Reich, 2025, S. 3).

- **Praktische Beispielszenarien zeigen, dass kontinuierliche Leistungsverbesserung und Feedbackintegration in Open-Source-Systemen notwendig sind, um wettbewerbsfähig zu bleiben, was jedoch zusätzliche Ressourcen für Monitoring, Evaluation und regelmäßige Modellaktualisierung erforderlich macht:** Die Stadt Nürnberg demonstriert im Rahmen des Projekts „Erfahrungsfeld-KI“, dass erfolgreiche KI-Integration nur durch gezielte Wissensvermittlung sowie den Aufbau spezifischer Fachkompetenzen bei den Mitarbeitenden nachhaltig möglich ist (Bahlinger et al., 2023, S. 2). Besonders im Kontext von KMU stellt die Kontrolle über Datenschutz, Sicherheit und betriebsspezifische Anpassungen einen erheblichen Vorteil, aber auch eine zusätzliche Belastung für die eigene IT-Abteilung dar (Dauser & Utomo, 2025, S. 5).

- Die Entwicklung hin zu immer größeren Modellen mit exponentiell wachsender Parameterzahl – GPT-1 mit 117 Millionen, GPT-2 mit 1,5 Milliarden, GPT-3 mit 175 Milliarden und GPT-4 mit rund 100 Billionen Parametern – verdeutlicht die Notwendigkeit, technologische und algorithmische Innovationen zur Effizienzsteigerung sowie Konzepte für nachhaltige Ressourcennutzung und Energiereduktion zu erforschen. Studien zeigen, dass etwa durch optimalen Einsatz effizienterer Hardware und Software sowie durch gezielte Anpassungen an spezifische Anwendungsfälle, wie im Marketing durch die Integration von KI-Assistenzsystemen bei Klarna, deutliche Produktivitätsgewinne und Reduktion von Ressourcenaufwand erzielt werden können (Saum, 2024, S. 11; Brüggenwirth et al., 2024, S. 11; Seeser-Reich, 2025, S. 3).

Passende Quellen:

- 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
- 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
- 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
- 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

5. Zukünftige Entwicklungen

5.1 Optimierungspotenziale

Zusammenfassung:
Diskussion der möglichen Optimierungspotenziale für Transformer-Modelle, einschließlich effizienterer Algorithmen, verbesserter Hardware-Nutzung und innovativer Techniken zur Reduktion des Ressourcenverbrauchs, um zukünftige Entwicklungen und Verbesserungen zu fördern.

Schwerpunkte:
Effizientere Nutzung und Weiterentwicklung des Self-Attention-Mechanismus zur Bewältigung langer Sequenzen: Der Self-Attention-Mechanismus der Transformer-Architektur ermöglicht eine flexible Modellierung von Wortbeziehungen innerhalb eines Texts, ist aber durch den exponentiell ansteigenden Ressourcenbedarf bei sehr langen Sequenzen limitiert. Neue Optimierungsansätze sind erforderlich, um die Kontextfenstergröße weiter zu erhöhen und gleichzeitig die Rechen- und Speicherressourcen zu senken, beispielsweise durch adaptives Sparsity oder segmentierte Verarbeitung. Die Analyse von Sequenzlängen, bei denen bereits für 10 Wörter 100.000 Einträge benötigt werden, demonstriert den dringenden Optimierungsbedarf in Bezug auf Speicher- und Effizienzsteigerung (Rosinski, 2022, S. 37).

Architekturinnovationen wie Enhanced Transformer und hybride Modelle führen nachweislich zu einer deutlichen Steigerung der Modellqualität: Das Einführen zusätzlicher architektonischer Komponenten, wie in der Enhanced-Transformer-Architektur mit vollständiger Layer-Normalisierung, gewichteten Residualverbindungen oder RL-basiertem Positionierungs-Encoding, resultiert in signifikant höheren BLEU-Scores, beispielsweise einer Verbesserung um bis zu 202,96 % gegenüber dem Standard-Transformer. Diese konkreten Fortschritte zeigen, wie gezielte Anpassungen auf Schichtebene oder durch alternative Verbindungsstrategien die Leistung für Aufgaben wie Übersetzung oder Dialoggenerierung markant steigern können (Moon et al., 2023, S. 1, S. 6).

Gezielte Kombination von Transformer-Modellen mit generativen adversarialen Netzen (GANs) ermöglicht eine Verbesserung der Chatbot-Dialogqualität: Studien belegen, dass hybride Ansätze, bei denen etwa ein Transformer als Generator mit einem cWGAN gekoppelt wird, zu höherwertigen, vielfältigeren und kontextuell passenderen Antworten führen. Der BLEU-4-Score im Chit-Chat-Datensatz verbessert sich deutlich, wenn ein vortrainiertes Transformer-Modell beim Training des Generators eingesetzt wird – ein Ansatz, der neue Wege für realistischere und robuster trainierte Chatbots aufzeigt (Esfandiari et al., 2023, S. 7, S. 9).

Spezifische Optimierung durch den gezielten Einsatz vortrainierter Embedding-Matrizen

und deren flexible Nutzung innerhalb der Transformer-Architektur: Die Auswahl zwischen selbst trainierten und extern vortrainierten Embedding-Matrizen, wie sie in modernen Systemen genutzt werden, bietet ein wesentliches Potenzial für die Anpassung an domänenspezifische Anforderungen und eine Steigerung der Modellgüte. Die Projektion der ursprünglichen Embedding-Vektoren in mehrere kleinere Vektorräume durch spezifische Gewichtsmatrizen erhöht die Effizienz und Präzision der Kontextrepräsentation bei gleichzeitiger Reduktion des Ressourcenverbrauchs (Krüger, 2021, S. 9, S. 15, S. 29).

Transferlernen und Zero-Shot-Inferenz als Schlüsselfaktoren für die Erweiterung der Multilingualität und Adaptierbarkeit von Chatbots: Durch die Fähigkeit von Transformer-Modellen, Transferlernen zu ermöglichen, und durch die Nutzung von Zero-Shot-Inferenz über Sprachpaare hinweg, können Systeme domänenübergreifend und für bisher ungesehene Sprachen oder Aufgaben eingesetzt werden. Dies schafft eine Grundlage für skalierbare, ressourceneffiziente und vielseitig einsetzbare Chatbot-Systeme, die gleichzeitig die Herausforderungen der Datenknappheit und der Heterogenität von Sprachdaten adressieren (Lakew et al., 2018, S. 1, S. 3, S. 7).

Passende Quellen:

- Esfandiari, N., Kiani, K., & Rastgoo, R. (2023). A Conditional Generative Chatbot using Transformer Model. Semnan University. https://arxiv.org/pdf/2306.02074
- 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
- 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
- Rosinski, W. (2022). Natural language processing. Hochschule Mittweida. https://www.staff.hs-mittweida.de/~rosinski/transformer/2022/presentation.pdf

6. Fazit