Literaturzusammenfassung

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

Bachelorstudium Informatik

Verwendete Quellen (24 Stück)

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

Link:
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

Relevante Kernergebnisse:
- Das Konzept wird als Zylinder mit einem weißen Kaninchen, einem schwarzen Kaninchen und einem Hohlraum dazwischen dargestellt (S. 1).
- Das weiße Kaninchen repräsentiert das natürliche Sprachverständnis oder die Kodierung (S. 1).
- Das schwarze Kaninchen steht für die natürliche Sprachgenerierung oder Dekodierung (S. 1).
- Der Hohlraum dazwischen wird als natürliche Sprachverarbeitung bezeichnet (S. 1).
- KI-Modelle (Claude und ChatGPT) haben die Analogie aufgegriffen und versucht, sie weiter auszuführen oder zu erklären (S. 1).
- Die Analogie erfasst die Grundkonzepte des Verständnisses, der Verarbeitung und der Generierung von natürlicher Sprache, die in diesen Modellen eine wichtige Rolle spielen (S. 1).

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

Link: https://opus4.kobv.de/opus4-ohm/files/1191/nbg_erfahrungsfeld_ohmdok_001.pdf

Relevante Kernergebnisse:
- Die Stadt Nürnberg ist sich des Fachkräftemangels bewusst und dass innovative Digitalisierungsansätze nur unter Einbezug von Methoden der KI zu integrieren sind, wenn Mitarbeitende über ein hinreichend dichtes Wissen verfügen. (S. 1)
- Zentraler Entwicklungsansatz ist der Aufbau eines eigenen Erfahrungshintergrunds zur Funktionsweise von KI. (S. 2)

- Mit dem Projekt „Erfahrungsfeld-KI“ und dem parallel initiierten „KI-Potenzial-Scanning” werden konkrete Fallstudien vorgestellt, Anwendungsbeispiele erklärt und erlebbar gemacht. (S. 2)
- Die Konzeptionsphase der Projekte begann mit gemeinsamen Workshops, in welchen die für das Vorhaben wesentlichen Abteilungen sondiert und ihre jeweiligen Wissensträger ermittelt wurden. (S. 3)
- Als Attribute für die Auswahl wurden unter anderem vorhandene Bürgerkontakte, derzeitige Kapazitäten der jeweiligen Abteilungen aber auch interne Faktoren herangezogen. (S. 3)
- Zu Beginn sollte das Erfahrungsfeld KI jedoch mindestens vier vollständig ausgearbeitete Stationen anbieten. (S. 11)

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

Link:
https://opus4.kobv.de/opus4-oth-regensburg/files/7129/Baumann_Sprachassistenten_proceedings.pdf#page=18

Relevante Kernergebnisse:
- AV-ASR übertrifft das akustische ASR in allen SNRs signifikant (S. 13)
- In verrauschteren Umgebungen zeigt das AV-ASR eine größere WER-Verbesserung im Vergleich zum akustischen ASR und zeigt die Rauschrobustheit des AV-ASR in verrauschten Umgebungen (S. 13)
- In einer Online-Umfrage gaben 45,7 % der Befragten an, dass sie ein fehlendes Vertrauen in die Technologie oder Sicherheit von Sprachassistenten haben (S. 30)
- 42 % der Teilnehmer zweifeln an der Benutzerfreundlichkeit und Akzeptanz von Sprachassistenten beim Personal (S. 30)
- Nur 18,5 % der Antworten betreffen nützliche Aufgaben für einen Sprachassistenten oder die Kosten (S. 30)
- Die Varianz im Interaktionsverhalten von Testnutzern zeigte, dass (1) frühzeitige Nutzerintegration und (2) Einschluss eines breiten Testnutzerspektrums für den Entwicklungs- und Optimierungsprozess entscheidend sind (S. 15)

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

Link: https://wts.de/wts.de/KI%20Studie/2024-04-11_KI-Folgestudie.pdf

Relevante Kernergebnisse:
- Der AI Index Report der Stanford University aus dem Jahr 2023 zeigt, dass generative Ansätze wie GAN (8%), NL Generation (12%) oder Transformer (11%) bisher nur begrenzt in ihre Unternehmenstätigkeiten integriert haben (S. 11).
- Eine KI erstmalig die Steuerfachangestellten-Prüfung mit knapp 54% bestanden hätte (S. 11).
- In einer Kurzstudie von Taxy.io zeigte sich, zu welchen Leistungen solche Modelle im Steuerkontext in der Lage sind (S. 11).
- Das Ziel: 80-90% der Sachverhalte werden über geführte Prozesse und KI abgebildet. Die restlichen 10% werden weiterhin manuell von qualitativ hochwertigen Fachkräften betreut (S. 31).
- Die automatisierte Verarbeitung von Texten ist eine der großen Stärken von Sprachmodellen (S. 35).
- Im Bereich der HR-Taxes kam KI praktisch kaum zum Einsatz. Am ehesten wurde hier in den letzten Jahren auf die Nutzung von OCR-Lösungen zurückgegriffen, um PDF-Dokumente auszulesen (S. 31).

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

Link:
https://www.vde.com/resource/blob/2359152/bc0c7b6d8464dc8e285618b35b11caa7/vde-position-paper-large-language-models-data.pdf

Relevante Kernergebnisse:
- OpenAIs ChatGPT, das speziell das GPT-4 LLM nutzt, verfügt über eine Nutzerbasis von 200 Millionen aktiven Nutzern weltweit und 1,5 Milliarden Besuchen pro Monat (S. 6).
- GPT-1 besteht aus 117 Millionen Netzwerkparametern, GPT-2 aus 1,5 Milliarden, GPT-3 aus 175 Milliarden und GPT-4 aus etwa 100 Billionen (S. 11).
- OpenGPT-X verwendet mehr als 2,5 Billionen Token für das Training von LLMs, bestehend aus etwa 80 % Webdaten und 20 % kuratierten Daten (S. 15).
- Die Filterung reduziert die Menge an Rohdaten um etwa 17 % und die Deduplizierung um

etwa 25 % (S. 15).
- OpenGPT-X hat die Verarbeitung von anfänglich zwei Sprachen (EN, DE) auf fünf Sprachen (EN, DE, FR, IT, ES) skaliert (S. 15).
- OpenAI plant Rechenzentren, die bis zu 5 GW Strom verbrauchen (S. 16).

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

Link: https://www.assets.dpunkt.de/leseproben/14157/Leseprobe.pdf

Relevante Kernergebnisse:
- GPT-4 Vision ermöglicht die Verarbeitung von Bildern als Eingabe in LLMs (S. 6)
- LLMs können Bilder mithilfe einer spezialisierten Transformer-Architektur namens Vision Transformer (ViT) verarbeiten (S. 6)
- N-Gramm-Modelle nutzen die Häufigkeit von Wörtern, um das nächste Wort vorherzusagen (S. 6)
- Transformer haben die Möglichkeit, den Kontext effektiv beizubehalten und zu codieren, was bei RNNs problematisch war (S. 6)
- Der Attention-Mechanismus ermöglicht direkte Verbindungen zwischen entfernten Elementen im Text, was eine signifikante Einschränkung älterer Modelle wie RNNs aufhebt (S. 7)
- Anfang 2024 wurden GPT-4 Turbo und GPT-40 von OpenAI veröffentlicht, die ein größeres Eingabekontextfenster von 128.000 Tokens aufweisen (S. 10)

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

Link: https://opus.bibliothek.uni-augsburg.de/opus4/files/113159/113159.pdf

Relevante Kernergebnisse:
- Transformer-Modelle ermöglichen eine tiefere Analyse von Texten und ein besseres Verständnis von Kontexten im Vergleich zu früheren Modellen (S. 3).
- Schlechte Datenqualität kann zu ungenauen und fehlerhaften Ergebnissen des maschinellen Lernens führen (S. 3).
- KI-Modelle haben oft Schwierigkeiten, Kontexte und feine Nuancen zu verstehen, was in

der Sprachverarbeitung problematisch sein kann (S. 3).
- Maschinelle Lernmodelle können dazu neigen, Fake News nicht zu erkennen und diese dadurch unbeabsichtigt weiter zu verbreiten (S. 3).
- SciSpace bietet eine Datenbank mit mehr als 200 Mio. Einträgen aktueller wissenschaftlicher Artikel (S. 8).
- Verschiedene KI-Tools bieten Forschenden neue Methoden zur Identifikation, Analyse und Vernetzung relevanter Literatur (S. 13).

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

Link: https://arxiv.org/pdf/2412.02975

Relevante Kernergebnisse:
- Für jede Konstante L benötigt jeder Decoder-only Transformer mit L Schichten eine polynomiale Modelldimension (n?(1)), um eine sequentielle Komposition von L Funktionen über eine Eingabe von n Token durchzuführen (S. 1).
- Als Konsequenz ergibt sich aus den Ergebnissen: (1) der erste Trade-off zwischen Tiefe und Breite für Multi-Layer-Transformer, der zeigt, dass die L-stufige Kompositionsaufgabe für L-Layer-Modelle exponentiell schwieriger ist als für (L + 1)-Layer-Modelle; (2) eine bedingungslose Trennung zwischen Encoder und Decoder, die eine schwierige Aufgabe für Decoder aufzeigt, die von einem exponentiell flacheren und kleineren Encoder gelöst werden kann; (3) ein nachweisbarer Vorteil von Chain-of-Thought, der eine Aufgabe aufzeigt, die mit Chain-of-Thought exponentiell einfacher wird (S. 1).
- Für jede Konstante L ≤ Õ(log log(n)) konnte ein L-Layer Decoder-only Transformer keine L-sequentielle Funktionskomposition lösen, wenn Hdp < n^(2-4L) ist (S. 2).
- Für jede Konstante L > 1 gibt es eine Aufgabe (a.k.a. L-sequentielle Funktionskomposition), so dass (1) ein (L + 1)-Layer Transformer die Aufgabe mit polylogarithmischer Anzahl von Parametern lösen könnte, während (2) jeder L-Layer Transformer eine polynomiale Anzahl von Parametern benötigt, um die Aufgabe zu lösen (S. 4).
- Für jede Konstante L > 1 gibt es eine Aufgabe (a.k.a. L-sequentielle Funktionskomposition), so dass (1) ein O(log(L))-Layer Transformer Encoder die Aufgabe mit polylogarithmischer Anzahl von Parametern lösen könnte, während (2) jeder L-Layer Transformer Decoder eine polynomiale Anzahl von Parametern benötigt, um die Aufgabe zu lösen (S. 4).
- Für jede Konstante L ≥ 1 gibt es eine Aufgabe (a.k.a. L-sequentielle Funktionskomposition), so dass (1) ein One-Layer Transformer mit CoT die Aufgabe mit polylogarithmischer Anzahl von Parametern lösen könnte, während (2) jeder L-Layer Transformer Decoder eine polynomiale Anzahl von Parametern benötigt, um die Aufgabe zu lösen (S. 4).

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

Link:
https://www.f-bb.de/fileadmin/Projekte/ZZBY/250204_ZZSUE_f-bb-online_KI-Experimentierraum.pdf

Relevante Kernergebnisse:
- KMU können durch eigene Chatbots die Datenschutz- und Sicherheitsmaßnahmen direkt kontrollieren, was für sensible Kundendaten entscheidend ist (S. 5).
- KI-Chatbots können KMU dabei helfen, wettbewerbsfähiger zu werden, ihre Betriebsabläufe zu optimieren und eine bessere Kundenzufriedenheit zu erreichen (S. 6).
- Regelmäßige Aktualisierung und die Einbeziehung von Feedback sind entscheidend, um die Leistung von Open-Source-Sprachmodellen langfristig zu verbessern (S. 10).
- Die Leistungsfähigkeit der entwickelten KI-Chatbots kann mit marktführenden KI-Systemen wie ChatGPT von OpenAI nicht mithalten (S. 11).
- Die Kosten für moderne LLM-basierte Chatbots liegen typischerweise zwischen 600 € und 2.000 € pro Monat (S. 17).
- Für die Anwendungsfälle der entwickelten Chatbots wären mindestens 64 GB RAM pro Server erforderlich (S. 17).

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

Link:
https://www.haw-landshut.de/static/ITZ/Bilder/Veranstaltungen/LLF/Beitraege_LL/2024/LL_4_2024_GenKI_Bild-Video_Ecker_Dotzauer.pdf

Relevante Kernergebnisse:
- Der große Aufschwung der ChatGPT-Technologie kam mit der GPT-3 Version, welche 2020 erschienen ist. (S. 2)
- DALL-E ist mit 12 Milliarden Parametern ausgestattet und darauf spezialisiert, visuelle Darstellungen aus reinen Textanweisungen zu generieren. (S. 7)
- DALL-E3 kann seit dem 6. Dezember 2023 kostenlos über den Bing Image Creator genutzt werden. (S. 7)
- Midjourney wurde im Jahr 2022 eingeführt und hat mit der Version 5.2 einen signifikanten Fortschritt erzielt. (S. 9)

- In Unternehmen ist eine deutliche Steigerung der Nutzung oder geplante Anwendung von KI erkennbar (S. 10)
- In der Radiologie ermöglicht KI bereits erhebliche Fortschritte, beispielsweise bei der automatischen Erkennung von Schlaganfällen in CT-Scans und der Erstellung von Prognosen bei Brustkrebspatienten zu deren Reaktion auf eine Chemo-Therapie. (S. 18)

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

Link: https://arxiv.org/pdf/2306.02074

Relevante Kernergebnisse:
- Das vorgeschlagene Modell zeigt eine bessere Leistung auf dem Chit-Chat-Datensatz (S. 7)
- Auf dem Cornell-Datensatz übertrifft das vorgeschlagene Modell alle anderen Modelle, da es durch Transformer reichhaltigere Merkmale extrahiert und die Datenverteilung durch GAN lernt (S. 7)
- Das vorgeschlagene Modell übertrifft alle Ansätze unter Verwendung der BLEU-4-Metrik im Chit-Chat-Datensatz, da es eine erweiterte Hybridmethode aus Transformer-Modell und cWGAN verwendet (S. 7)
- Bei der besten Anstrengung werden 200 Zyklen verwendet, um den Transformer als vortrainiertes Modell des Generators zu trainieren (S. 9)
- Das Modell, das nur mit Transformer trainiert wird, ohne das gegnerische Lernen einzubeziehen, die geringste Leistung aufweist (S. 9)
- Das vortrainierte Modell verbessert die Leistung des Chatbots in allen Metriken des Cornell-Datensatzes (S. 9)

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

Link:
https://library.oapen.org/bitstream/handle/20.500.12657/99902/9783658447816.pdf?sequence=1#page=74

Relevante Kernergebnisse:
- Machine Learning hat seit der Jahrtausendwende beachtliche Erfolge erzielt (S. 11).
- Künstliche Intelligenz wird oft mit Machine Learning gleichgesetzt, obwohl sie mehr umfasst (S. 11).
- Transformer-Architekturen werden zur Generierung sprachlich kohärenter Texte verwendet

(S. 13).
- Hybride KI-Anwendungen sind bereits in der Praxis angekommen (S. 8).
- Die Kapitel sind in vier Gruppen gegliedert: Systeme, bei denen das Lernen dem Schlussfolgern hilft; Systeme, bei denen das Schlussfolgern dem Lernen hilft; Systeme, bei denen Lernen und Schlussfolgern eng gekoppelt sind; und Systeme, bei denen sie nur lose gekoppelt sind (S. 6).
- Die Anzahl der in einem Bild generierten Finger ist stellenweise inkohärent (S. 13).

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

Link:
https://www.irjmets.com/uploadedfiles/paper//issue_4_april_2025/73761/final/fin_irjmets1745500376.pdf

Relevante Kernergebnisse:
- Im IMDB-Datensatz erreichte Transformer eine Genauigkeit von 90,3 % im Vergleich zu 87,1 % für BILSTM (S. 2).
- Bei der Nachrichtenklassifizierung (AG News) erreichte Transformer eine Genauigkeit von 93,6 % im Vergleich zu 91,2 % für BiLSTM (S. 2).
- Im WMT'14 EN-DE-Datensatz erzielte Transformer einen BLEU-Score von 28,9 im Vergleich zu 25,8 für BiLSTM und trainierte schneller (S. 2).
- Transformer übertraf RNNs in Bezug auf Genauigkeit und Trainingseffizienz über alle Aufgaben hinweg (S. 2).
- RNNs bleiben für Umgebungen mit geringen Ressourcen relevant (S. 2).
- Die Transformer-Modellkonfiguration umfasste 6 Schichten, eine Hidden Size von 512, 8 Attention Heads und einen Dropout von 0,1 (S. 3).

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

Link:
https://www.trans-kom.eu/bd14nr02/trans-kom_14_02_05_Krueger_NMUe.20211202.pdf

Relevante Kernergebnisse:

- Im professionellen Fachübersetzen planten 78% aller im Rahmen der European Language Industry Survey 2020 befragten Unternehmen, MÜ-Systeme oder Post-Editing-Workflows in ihre Arbeitsprozesse zu integrieren oder den Einsatz dieser Systeme und Workflows auszuweiten (S. 1).
- Bis zur Einführung der Transformer-Architektur im Jahr 2017 basierten leistungsstarke NMÜ-Systeme meist auf sogenannten rekurrenten neuronalen Netzen (RNN) (S. 2).
- Mit einer solchen RNN-Architektur sind gewisse Nachteile verbunden. So erschwert beispielsweise die sequenzielle Verarbeitung der Input-Sequenz die korrekte Identifizierung von Wortdependenzen über längere Distanzen hinweg (S. 2).
- Die Embedding-Matrix kann speziell für eine Transformer-Implementation trainiert werden oder es können externe, bereits vortrainierte Embedding-Matrizen genutzt werden (siehe z. B. das für die Vektor-berechnung in Abb. 6 genutzte Word-Embedding-Modell) (S. 9).
- Die ursprünglichen Word-Embedding-Vektoren haben eine Dimensionalität von 512 (vgl. Abschnitt 3.1.1), die für diese Embedding-Vektoren erzeugten Query-, Key- und Value-Vektoren haben dagegen nur eine Dimensionalität von 64 (S. 15).
- Die ursprünglichen Embedding-Vektoren werden in jeweils acht separate Query-, Key- und Value-Vektoren aufgespalten, in der linearen Schicht mit den einzelnen Gewichtsmatrizen multipliziert und dadurch in acht verschiedene 64-dimensionale Vektorräume projiziert (S. 29).

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

Link: https://aclanthology.org/C18-1054.pdf

Relevante Kernergebnisse:
- Multilinguale NMT zeigte wettbewerbsfähige Leistung gegenüber rein zweisprachigen Systemen. (S. 1)
- In Umgebungen mit geringen Ressourcen erwies es sich als effektiv und effizient, dank des gemeinsamen Darstellungsraums, der über Sprachen hinweg erzwungen wird und eine Art Transferlernen induziert. (S. 1)
- Multilinguale NMT ermöglicht sogenannte Zero-Shot-Inferenz über Sprachpaare, die zur Trainingszeit noch nie gesehen wurden. (S. 1)
- Die Transformer-Architektur arbeitet mit einem Selbstaufmerksamkeitsmechanismus (Intra-Attention), wodurch alle rekurrenten Operationen, die im vorherigen Ansatz gefunden wurden, entfernt werden. (S. 3)
- Die Ergebnisse zeigen, dass Transformer in allen drei Modellvarianten besser abschneidet. (S. 6)
- Die Transformer-Architektur zeigt eine definitiv höhere Qualität als die RNN-Architektur und bestätigt die Fähigkeit des Ansatzes, ungesehene Richtungen abzuleiten. (S. 7)

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

Link:
https://www.bitkom.org/sites/main/files/2023-06/BitkomLeitfadenGrosse-Sprachmodelle.pdf

Relevante Kernergebnisse:
- ChatGPT hat seit seiner Veröffentlichung im November 2022 viel Aufmerksamkeit gewonnen und wird von Millionen von Nutzerinnen und Nutzern angewendet (S. 4).
- Transformer-Architektur wird bereits seit über fünf Jahren von Entwicklerinnen und Entwicklern maschineller Übersetzung genutzt (S. 5).
- Als zum Beispiel ChatGPT auf der Grundlage von GPT 3.5 aufgefordert wurde, die Fläche eines Dreiecks mit einer Höhe von 2,5 Metern und einer Basis von 3,7 Metern zu berechnen, gab es folgende Antwort: Fläche = 4,625 Quadratmeter (S. 6).
- Das gesammelte Feedback führte zu den erheblichen Verbesserungen in der Version GPT-3.5, durch die die Anwendung ChatGPT in ihrer ersten Veröffentlichung so viel Aufmerksamkeit erregte (S. 8).
- 80 Prozent der Unternehmen haben ihren Hauptsitz in Deutschland, jeweils 8 Prozent kommen aus Europa und den USA, 4 Prozent aus anderen Regionen (S. 17).
- Bitkom vertritt mehr als 2.000 Mitgliedsunternehmen aus der digitalen Wirtschaft (S. 17).

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

Link:
https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/teaching/archive/ws_1819/vl_maschsprache/07_classification.pdf

Relevante Kernergebnisse:
- Unterschiede bei der Verwendung verschiedener Klassifikationsmethoden auf denselben Daten/Darstellungen sind oft erstaunlich gering (S. 19).
- KNN ist sehr einfach und effektiv, aber langsam (S. 20).
- Naive Bayes ist ein effizientes Lernverfahren mit einem speichereffizienten Modell (O(|K|*|C|) Speicherplatz) (S. 32).
- In der Regel übertrifft ME NB (S. 46).
- Schätzungsweise >95% aller E-Mails sind Spam (S. 82).
- Bei der Spam-Erkennung scheinen das Entfernen von Stoppwörtern und das Stemming (gemischte Berichte) zu helfen (S. 83).

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

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

Relevante Kernergebnisse:
- Lediglich neun Artikel (17%) betrachten den Einsatz von Chatbots am Büroarbeitsplatz bzw. in einem büroarbeitsplatznahen Einsatz. (S. 20)
- Dies resultierte in 43 relevante Artikel (83%), die Aspekte betrachten, die auf den Einsatz am digitalen Büroarbeitsplatz übertragbar sind. (S. 20)
- Die Verteilung nach Sprachen zeigt, dass überwiegend englischsprachige Artikel (n=37; 71 %) identifiziert wurden. (S. 21)
- Nur 15 Artikel (29%) sind auf Deutsch veröffentlicht. (S. 21)
- Die Betrachtung der Publikationsart ergibt, dass hauptsächliche Konferenzbeiträge (n=22, 42%) das Forschungsthema aufgreifen, dicht gefolgt von wissenschaftlichen und praxisbezogenen – Fachzeitschriften (n=21; 41%). (S. 21)
- Circa 23 Artikel (44%) stellen praxisbezogene Veröffentlichungen dar. (S. 29)
- Die meisten Artikel (n=29; 56%) richten sich aber an ein wissenschaftliches Publikum. (S. 29)

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

Link: https://arxiv.org/pdf/2310.10930

Relevante Kernergebnisse:
- Der Enhanced Transformer erzielt einen um 202,96 % höheren BLEU-Score im Vergleich zum ursprünglichen Transformer mit dem Übersetzungsdatensatz (S. 1)
- Der durchschnittliche BLEU-Score des Originalmodells beträgt 18,59 (S. 6)
- Der durchschnittliche BLEU-Score des Modells mit vollständiger Layer-Normalisierung beträgt 42,31, was einer Verbesserung von 127,60 % gegenüber dem Originalmodell entspricht (S. 6)

- Der durchschnittliche BLEU-Score des Modells mit gewichteter Residualverbindung beträgt 29,70, was einer Verbesserung von 59,76 % gegenüber dem Originalmodell entspricht (S. 6)
- Der durchschnittliche BLEU-Score des Modells mit Positionierungs-Encoding unter Nutzung von RL beträgt 28,58, was einer Verbesserung von 53,74 % gegenüber dem Originalmodell entspricht (S. 6)
- Der durchschnittliche BLEU-Score des Modells mit Zero-Masked-Self-Attention beträgt 25,16, was einer Verbesserung von 35,34 % entspricht (S. 6)

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

Link: https://aclanthology.org/2024.nlp4dh-1.37.pdf

Relevante Kernergebnisse:
- Naive Bayes erreichte eine maximale Genauigkeit von 64% bei der Klassifizierung von Hindi-Gedichten. (S. 2)
- Das SVM-Modell erreichte die höchste Genauigkeit (75%) bei einer Sentiment-Klassifizierungsaufgabe für die Mizo-Sprache. (S. 2)
- Die Theragatha besteht aus 1.288 Versen, die auf 21 Kapitel verteilt sind. (S. 3)
- Die Therigatha enthält 524 Verse, die auf 16 Kapitel verteilt sind. (S. 3)
- Bei Experimenten mit dem Devanagari-Skript schnitten Transformer-basierte Modelle im Vergleich zu ihren Pendants im römischen Skript schlechter ab. (S. 4)
- Die Anzahl der eindeutigen Token der Tokenizer im Test-Subset korrelierte stark mit den Klassifizierungsergebnissen. (S. 4)

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

Link: https://www.staff.hs-mittweida.de/~rosinski/transformer/2022/presentation.pdf

Relevante Kernergebnisse:
- Google entwickelte 2013 word2vec-Modelle mit 6 Milliarden Trainingswörtern, die erstmals konkurrenzfähige Ergebnisse im Vergleich zu klassischen Modellen zeigten (S. 32).
- AI2 und die University of Washington entwickelten 2017 das ELMo-Modell, ein kontextuelles Einbettungsmodell mit bidirektionalen 3-Schicht-LSTM und 1 Milliarde

Trainingswörtern (S. 32).
- Google entwickelte 2018 Einbettungsmodelle mittels Transformer, eine neue NN-Architektur basierend auf Aufmerksamkeit, die eine bessere Effizienz für das Training von Modellen im großen Maßstab ermöglicht (S. 32).
- Zu den ersten Modellen, die auf Transformer basieren, gehören GPT (OpenAPI) und BERT (Google) mit bidirektionalen Transformer (S. 32).
- Aktuell modernste Einbettungsmodelle sind GPT3 von OpenAI mit 170 Milliarden Parametern und GShard von Google mit 600 Milliarden Parametern (S. 32).
- Die Analyse von Sequenzlängen von 10 Wörtern benötigt 100.000 Einträge (S. 37).

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

Link: https://arxiv.org/pdf/2306.02896

Relevante Kernergebnisse:
- Transformer-Modelle zeigen empirisch Verzerrungen gegenüber bestimmten algorithmischen Primitiven, wenn sie mit gradientenbasierten Lernalgorithmen trainiert werden (S. 1).
- Die Kontextlänge von GPT-4 beträgt bis zu N = 32000 (S. 2).
- Es existiert eine Self-Attention-Einheit mit einer m-dimensionalen Einbettung, die qSA genau dann approximiert, wenn m ≥ q (S. 3).
- Eine einzelne Schicht von Standard-Multi-Head-Self-Attention kann Match3 nicht berechnen, es sei denn, die Anzahl der Köpfe H oder die Einbettungsdimension m wächst polynomial in N (S. 3).
- Für ausreichend großes q, N ≥ 2q + 1 und d' ≥ 1 existiert eine universelle Konstante c, so dass, wenn mp < cq, keine Funktion f existiert, die qSA approximiert (S. 9).
- Jede deterministische Protokoll zur Berechnung von DISJ(a,b) benötigt mindestens n Kommunikationsrunden (S. 10).

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

Link:
https://www.unternehmerinnenforum-niederrhein.de/wp-content/uploads/2024/06/Saum_KI-im-Marketing_Skript.pdf

Relevante Kernergebnisse:

- China hat das Ziel, bis 2030 weltweit führend im Bereich KI zu sein (S. 3).
- Hugging Face, eine Plattform für Open-Source-Modelle und Datensätze, hat bereits eine Bewertung von 2 Milliarden US-Dollar erreicht (S. 3).
- Eine Studie von BCG und der Harvard Business School aus dem Jahr 2023 ergab eine durchschnittliche Effizienzsteigerung von 12,2% und eine Zeitersparnis von 25,1% bei den GPT-4-Nutzern (S. 11).
- Dieselbe Studie zeigte, dass 40% der GPT-4-Nutzer qualitativ bessere Ergebnisse lieferten (S. 11).
- Ein KI-Assistent bei Klarna bearbeitete im ersten Monat 2,3 Mio. Unterhaltungen, was 2/3 aller Anfragen entsprach (S. 13).
- Bei Klarna führte der KI-Assistent zu einer schnelleren Bearbeitung (2 statt 11 Minuten) und besseren Lösungen (-25% wiederholende Anfragen) (S. 13).

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

Link: https://forum.dguv.de/issues/RZ_S016-018_1.04_Chatbots.pdf

Relevante Kernergebnisse:
- 94 Prozent der von KI erstellten Hausarbeiten im Fach Psychologie wurden nicht als solche erkannt und im Schnitt sogar bessere Noten erzielt (S. 1).
- SmarterChild führte innerhalb eines Jahres mehr als neun Millionen Konversationen (S. 1).
- ChatGPT verzeichnet wöchentlich 100 Millionen aktive Nutzer (S. 2).
- LLMs können den Gesprächskontext nur für eine bestimmte Anzahl von Wörtern speichern (S. 2).
- Moderne Chatbots benötigen riesige Datenmengen und leistungsfähige Computer für Training und Betrieb (S. 3).
- Chatbots können überzeugend klingende, aber falsche Informationen generieren, ein Phänomen, das als „Halluzination" bezeichnet wird (S. 3).