# Pegelrechnung mit ChatGPT - Hot or not?

Tom Ehrig[1], Martin Dannemann[2], Niels Modler[3]

[1] *TU Dresden, Institut für Leichtbau und Kunststofftechnik, 01307 Dresden, E-Mail: tom.ehrig@tu-dresden.de*
[2] *Westsächsische Hochschule Zwickau, Institut für Energie und Verkehr, 08012 Zwickau, E-Mail: martin.dannemann@fh-zwickau.de*
[3] *TU Dresden, Institut für Leichtbau und Kunststofftechnik, 01307 Dresden, E-Mail: niels.modler@tu-dresden.de*

## Einleitung und Motivation

Mit der zunehmenden Verbreitung von *Large Language Models* (LLM), wie ChatGPT, welches im November 2022 der Öffentlichkeit zugänglich gemacht wurde, hat eine neue Ära der künstlichen Intelligenz begonnen. Mit intuitiven Benutzeroberflächen haben diese Modelle in kürzester Zeit ein breites Anwendungsspektrum gefunden, das von der Unterstützung beim Verfassen von E-Mails, Briefen oder kreativen Texten bis hin zur Lösung komplexer Fragestellungen in zahlreichen Fachgebieten reicht. Trotz der beeindruckenden Fähigkeiten, komplexe Prüfungsfragen aus verschiedenen Wissensgebieten korrekt zu beantworten [1][2], hat die Nutzung von ChatGPT nicht nur in der akademischen Lehre und Forschung Fragen bezüglich seiner Zuverlässigkeit und Genauigkeit aufgeworfen. Insbesondere im Kontext der Akustik – einem Fachbereich, der präzise Terminologie und komplexe Berechnungen erfordert – sind die Leistungen von ChatGPT Gegenstand intensiver Diskussionen geworden.

LLMs verarbeiten Eingabedaten und -prompts zu Texten und berechnen mit Hilfe von sogenannten *tokens* eine Wahrscheinlichkeit dafür, welche Phrase oder Wort in einem gegebenen Kontext folgt. Dabei offenbaren LLMs ab einer gewissen Anzahl von Trainingsparametern emergente Fähigkeiten, d. h. sie zeigen Eigenschaften, auf welche sie nicht trainiert wurden und die weit über ihre ursprüngliche Funktionalität hinausgehen [3]. So kann ChatGPT Rechenaufgaben richtig lösen, obwohl keine spezielle Mathematik- oder Rechenlogik implementiert wurde. Die Trainingsparameter dieser Modelle werden aus einer Vielzahl öffentlich zugänglicher online Quellen extrahiert, darunter Buch- und Artikeldatenbanken, Wikipedia- und Nachrichtenartikel, Forenbeiträge sowie Inhalte sozialer Medien.

Im akustischen Kontext kursieren jedoch zahlreiche Falschaussagen im Internet und halten sich dort hartnäckig (eine beispielhafte Sammlung hierzu findet sich in [4]). Dies betrifft sowohl die Verwendung von Fachbegriffen wie Schalldruck- oder Schallleistungspegel als auch Aufgaben zur Pegelrechnung. Eine weitere Herausforderung stellt die oft verkürzte Schreibweise dar (z. B. Schallpegel, statt Schalldruckpegel, Schallleistungspegel oder Schallintensitätspegel), die nur im Kontext eindeutig ist.

Diese Problematik führt zur zentralen These dieser Arbeit: Das Sprachmodell ChatGPT kann (komplexe) Aufgaben im akustischen Kontext entweder gar nicht oder nur mit einer erhöhten Fehlerquote lösen. Ziel der durchgeführten Untersuchungen war es, diese These mit Hilfe mehrerer Prompts zum Themenfeld akustischer Pegelrechnung zu prüfen und die Antworten der frei verfügbaren GPT-Version 3.5 mit denen der neueren Version GPT-4.0 zu vergleichen.

## Methodik

### Vorbereitung

Zunächst wurde ein Fragenkatalog erstellt, der sich aus Fragen aus Fachbüchern, Übungen und Praktika zusammensetzte. Diese Fragen deckten ein breites Spektrum akustischer Themen ab und wurden in folgende Kategorien geclustert: (1) einfache Pegeladditionen; (2) Textaufgaben zur Berechnung von Korrekturfaktoren; (3) Berechnungen von Schalldruckpegeln aus gegebenen Schallleistungspegeln; (4) komplexe Textaufgaben zur Bestimmung von Gesamtschallleistungspegeln. Die Ergebnisse werden im Folgenden an jeweils einer exemplarischen Beispielfrage aus den genannten vier Kategorien diskutiert.

### Durchführung

Die Untersuchungen wurden im September 2023 mit der Browserversion der „ChatGPT September 25 Version“ durchgeführt. Dabei wurden getrennte OpenAI-Konten für die Nutzung von GPT-3.5 und GPT-4.0 verwendet [5]. Für jede Frage wurde ein neuer Chat gestartet und der Eingabeprompt abgeschickt. Mit der *regenerate* Schaltfläche wurden für jede Frage 20 Antworten generiert und alle Antworten gespeichert. Dabei wurden keine Plugins wie z.B. *Calculator* genutzt, die Antworten nicht mittels „Daumen hoch“ oder „Daumen runter“ bewertet und keine Rückfragen gestellt. Die Durchführung der Untersuchung erfolgte über mehrere Tage verteilt, da die Anzahl an Anfragen pro Stunde begrenzt war.

Die Bewertung der Antworten erfolgte nur mit richtig oder falsch, ohne die Berücksichtigung von „Folgefehlern“. Da erste Vorversuche gezeigt haben, das ChatGPT häufig lange Textantworten mit Herleitung und oft sogar widersprüchlichen Aussagen innerhalb der Antwort liefert, die eine Bewertung erschweren, wurden die Prompts weiter präzisiert.

## Ergebnisse

### Exemplarisches Beispiel 1: Pegeladdition

Als exemplarisches Beispiel für eine gestellte Aufgabe zur Pegeladdition ist in Abb. 1a der Eingabeprompt sowie die prozentuale Häufigkeit, mit der die beiden GPT-Versionen 3.5 und 4.0 die Aufgabe richtig beantwortet haben, dargestellt. Während GPT-4.0 die Frage immer richtig beantwortet hat, liefert GPT-3.5 eine breite Spanne falscher Antworten, von 8 dB bis zu unrealistischen 8.000.000 dB. Eine genaue Analyse der Lösungswege zeigte dabei, dass häufig eine Vermischung der im Deutschen und Englischen unterschiedlichen Dezimaltrennzeichen sowie eine fehlerhafte Berechnung des Logarithmus die Ursache waren.

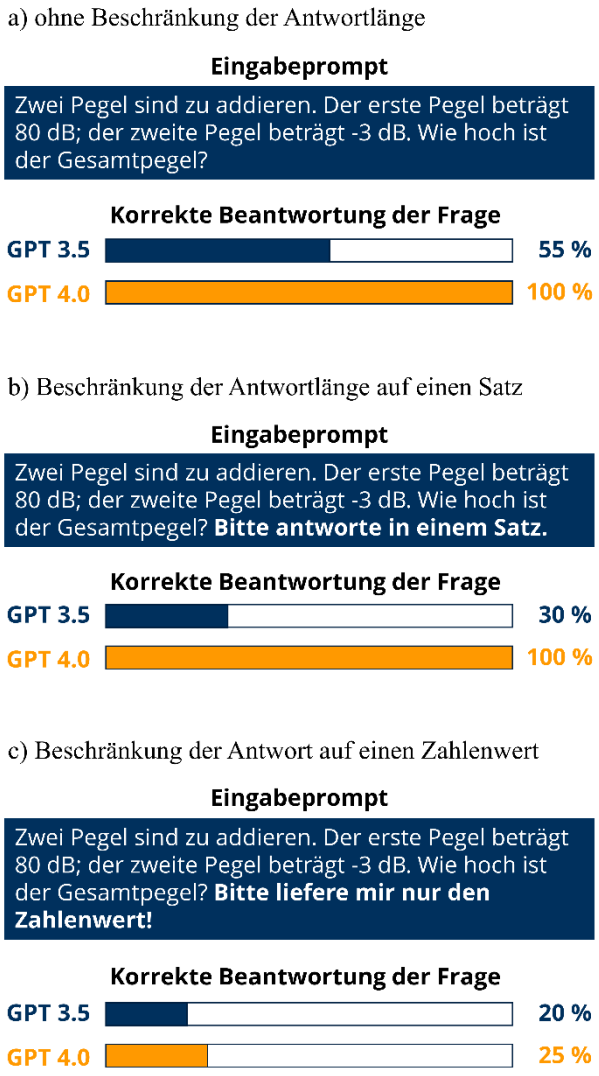


**Abbildung 1:** Aufgabe zur Pegeladdition; Eingabeprompt und prozentuale Häufigkeit der richtigen Antworten.

Da beide Modellversionen sehr ausführliche Antworten lieferten, wurde der Eingabeprompt um den Satz „Bitte antworte in einem Satz.“ ergänzt (Abb. 1b). GPT-4.0 antwortet häufig trotzdem mit mehr als einem Satz, liefert jedoch weiterhin immer das korrekte Ergebnis. Bei GPT-3.5 erhöht diese Beschränkung die Fehlerquote.

Um die häufigen Relativierungen in den Antworten zu vermeiden und präzisere Ausgaben zu erzwingen, wurde der Eingabeprompt noch einmal modifiziert („Bitte liefere mir nur den Zahlenwert!“, Abb. 1c). Dies führt zu dem überraschenden Ergebnis, dass sich die Fehlerquote bei GPT-3.5 noch weiter erhöht und nun auch GPT-4.0 nur noch zu 25 % die Frage korrekt beantwortet. Dabei ist interessant, dass in allen Fällen, bei denen GPT-4.0 die Frage korrekt beantwortet hat, die Antwort nicht als Zahlenwert, sondern in einem Satz erfolgte. In allen anderen Fällen wurde zwar nur mit einem Zahlenwert geantwortet, jedoch war dies immer falsch.

Zusammenfassend kann festgehalten werden, dass die Aufforderung zur Beschränkung auf einen reinen Zahlenwert zu einer höheren Fehlerquote führte. Diese Beobachtung unterstreicht die Sensitivität von ChatGPT auf die Formulierung der Eingabeprompts.

### Exemplarisches Beispiel 2: Korrekturfaktorberechnung

Auch bei der Berechnung eines Korrekturfaktors zeigt sich ein signifikanter Unterschied zwischen den beiden Modellvarianten (Abb. 2a).

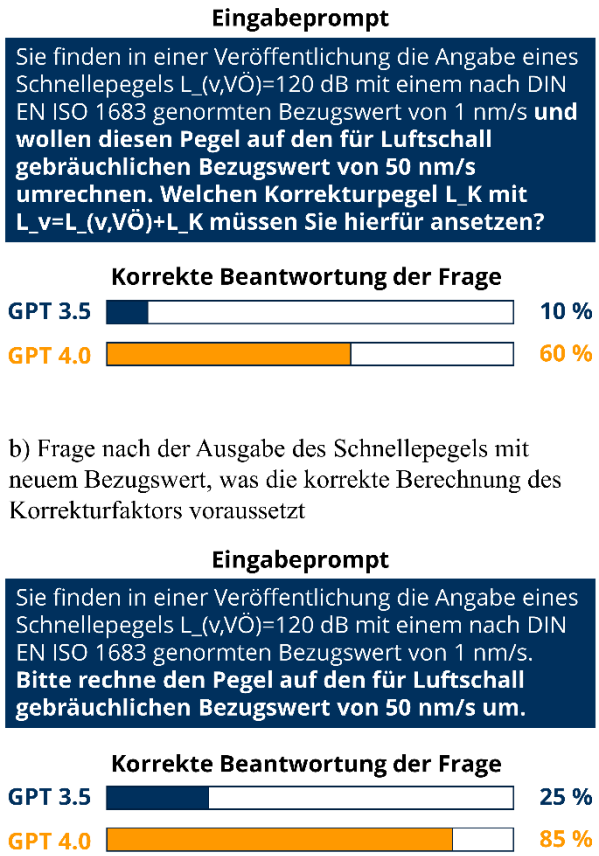


**Abbildung 2:** Aufgabe zur Korrekturfaktorberechnung; Eingabeprompt und prozentuale Häufigkeit der richtigen Antworten.

Dabei fällt auf, dass zwar oft der Betrag des Korrekturfaktors korrekt berechnet wird, jedoch ChatGPT das falsche

Vorzeichen ausgibt (GPT-3.5: 67 % und GPT-4.0: 50 % der falschen Antworten). Weiterhin zeigt die detaillierte Auswertung der Antworten, dass beide Modellversionen Schwierigkeiten hatten, zwischen Energie- und Feldgrößen (Vorfaktor 10 bzw. 20) zu unterscheiden, was in der akustischen Pegelrechnung eine kritische Rolle spielt. Nur sehr selten werden auch offensichtlich abwegige Ergebnisse, wie ein Korrekturfaktor von -680 dB, kritisch hinterfragt.

Obwohl im Eingabeprompt nicht danach gefragt wurde, berechnet ChatGPT fast immer auch den Schnellepegel für den neuen Bezugswert und gibt diesen mit aus. Aus diesem Grund wurde der zweite Teil des Eingabeprompts im nächsten Schritt noch einmal modifiziert (Abb. 2b). Beide Modellvarianten berechnen nun häufiger das korrekte Ergebnis. Überraschenderweise wurde nun auch der Korrekturpegel häufiger richtig berechnet und ausgegeben. Dies unterstreicht noch einmal, wie sensitiv ChatGPT auf Veränderungen des Eingabeprompts reagiert, auch wenn diese inhaltlich sehr ähnlich sind. Für den Nutzer ist dies ein Problem, da dieser kaum abschätzen kann, welches der „richtige Eingabeprompt“ ist.

**Exemplarisches Beispiel 3: Berechnung des Schalldruckpegels aus einem geg. Schallleistungspegel**

Bei dieser Aufgabe sollte für gegebene Randbedingungen aus einem Schallleistungspegel der Schalldruckpegel berechnet werden (Abb. 3). Obwohl im Eingabeprompt explizit angegeben, wurde in vielen Fällen mit der Randbedingung einer Vollkugelhüllfläche gerechnet (GPT-3.5: 35 % und GPT-4.0: 81 % der falschen Antworten). Dabei weißt ChatGPT oft sogar in den Antworten darauf hin, dass die Berechnung für eine Halbkugelhüllfläche erfolgen soll („*hier als Halbkugel-Hüllfläche beschrieben*“, „*In Ihrem Fall wird der Schall in Form einer Halbkugel ausbreiten*“ etc.), nutzt dann aber doch die Formel für den Vollkugelansatz. Es scheint somit in beiden Modellvarianten einen deutlichen Bias zum Vollkugelansatz zu geben.

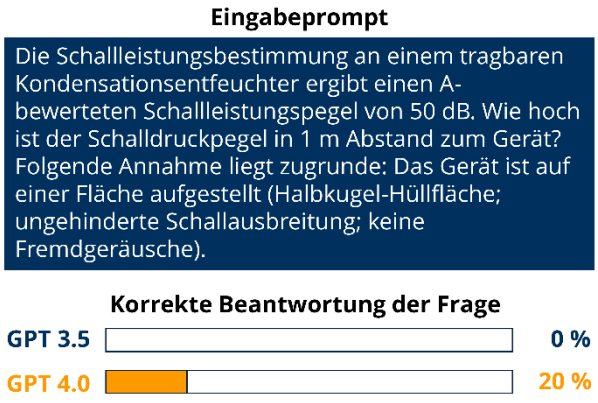


**Abbildung 3:** Aufgabe zur Berechnung des Schalldruckpegels aus einem geg. Schallleistungspegels; Eingabeprompt und prozentuale Häufigkeit der richtigen Antworten.

GPT-3.5 nimmt zudem zu 60 % immer die gleiche, falsche Formel zur Berechnung des Schalldruckpegels. Auffallend ist, dass GPT-4.0 diesen Fehler nicht macht. Weitere Fehler treten auch hier bei der Berechnung des Logarithmus sowie beim Umstellen der Gleichung auf.

**Exemplarisches Beispiel 4: komplexe Textaufgabe zur Bestimmung eines Gesamtschallleistungspegels**

Mit dieser Aufgabe sollte überprüft werden, inwieweit ChatGPT auch komplexe Textaufgaben mit einer sehr praxisnahen Aufgabenstellung analysieren und lösen kann (Abb. 4). Die Beantwortung der Frage wurde als korrekt gewertet, wenn die finale Aussage war: Ja, es geht (unabhängig von der konkreten Angabe, wieviel Reduktion notwendig ist).

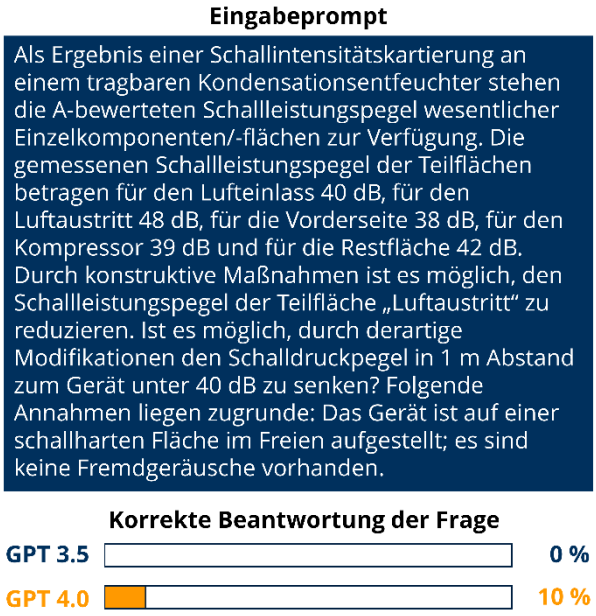


**Abbildung 4:** Komplexe Textaufgabe zur Bestimmung eines Gesamtschallleistungspegels; Eingabeprompt und prozentuale Häufigkeit der richtigen Antworten.

GPT-3.5 beschreibt in einigen Antworten einzelne „gute Lösungsansätze“, widerspricht sich dann aber in den folgenden Sätzen häufig selbst und konnte die Aufgabe nie korrekt lösen.
Im Gegensatz dazu beschreibt GPT-4.0 meist ein korrektes Vorgehen: zunächst die Berechnung des Gesamtschallleistungspegels aus den Schallleistungspegeln der Teilflächen und anschließend die Umrechnung auf den Schalldruckpegel. An dieser Stelle treten dann jedoch häufig Fehler auf oder ChatGPT „weiß nicht weiter“. Zum Beispiel wird auch kein einziges Mal der naheliegende Ansatz vorgeschlagen, den 48 dB-Pegel am Luftaustritt bei der Pegeladdition des Gesamtschallleistungspegels einfach wegzulassen (sprich eine fiktive maximale Reduktion anzunehmen), um die Antwort auf die gestellte Frage zu bekommen. Stattdessen werden Antworten wie „*bitte konsultieren Sie einen Akustiker*“, „*Sie müssen die Maßnahmen durchführen und messen*“, „*schwierig*“, „*unwahrscheinlich*“ etc. gegeben (Abb. 5).

## Zusammenfassung

Ziel der durchgeführten Untersuchungen war es, die These, das ChatGPT (komplexe) Aufgaben im akustischen Kontext

gar nicht oder nur mit einer erhöhten Fehlerquote lösen kann mit Hilfe mehrerer Prompts zur akustischer Pegelrechnung zu prüfen und die Antworten von ChatGPT in der Version GPT-3.5 mit denen der neueren Version GPT-4.0 zu vergleichen. Dabei konnte gezeigt werden, dass es selbst zu einer Fragestellung starke Unterschiede sowohl zwischen den unterschiedlichen GPT-Versionen als auch bei leicht veränderten, aber inhaltlich deckungsgleichen, Eingabeprompts kommt. So führt bspw. die Beschränkung der Ausgabe auf das Ergebnis („Bitte liefere mir nur den Zahlenwert!“) wesentlich häufiger zu Fehlern als die Ausgabe in Form eines Satz („Antworte in einem Satz!“). Diese Beobachtung unterstreicht die Sensitivität von ChatGPT auf die Formulierung der Anfragen und die Notwendigkeit einer sorgfältigen Gestaltung der Prompts, um korrekte Ergebnisse zu erzielen.

Die Ergebnisse verdeutlichen die Potenziale, aber auch die Grenzen von ChatGPT und ähnlichen Sprachmodellen in akademischen und technischen Anwendungsbereichen. Während GPT-4.0 eine signifikante Verbesserung gegenüber GPT-3.5 darstellt, offenbaren die Ergebnisse auch, dass die heutigen KI-Modelle häufig Fehler machen, insbesondere wenn die Anfragen nicht präzise formuliert sind oder die Aufgabenstellung besondere Anforderungen an die Verarbeitung und Interpretation der Daten stellt. Dabei werden oft richtige Ansätze gewählt, jedoch folgen dann z. T. widersprüchliche Aussagen oder die Sprachmodelle „verrechnen“ sich schlicht. Dabei erfolgt die Addition und Subtraktion meist fehlerfrei. Berechnungen mit Potenzen oder Logarithmus sind jedoch stark fehleranfällig, ebenso Aufgaben, bei denen das korrekte Vorzeichen eine Rolle spielt. Weitere häufige Fehlerquellen sind falsch interpretierte Dezimaltrennzeichnen und speziell im akustischen Kontext die richtige Unterscheidung zwischen Energie- und Feldgrößen (Vorfaktor 10 bzw. 20).

Die fortlaufende Entwicklung von KI-Technologien und Sprachmodellen verspricht weitere Verbesserungen in der Zukunft. Es ist zu erwarten, dass künftige Versionen von ChatGPT und anderen Sprachmodellen eine noch höhere Genauigkeit und Zuverlässigkeit in der Bearbeitung spezifischer, auch technisch komplexer Fragestellungen erreichen werden. Es empfiehlt sich, die Entwicklung dieser Technologien kontinuierlich zu beobachten und ihre Anwendung kritisch zu reflektieren. Anwender sollten sich der Grenzen der aktuellen Modelle bewusst sein und die Ergebnisse kritisch prüfen, insbesondere bei der Anwendung in Fachgebieten, die präzise und zuverlässige Daten erfordern. Die sorgfältige Formulierung von Anfragen und die kritische Bewertung der Antworten bleiben essentielle Kompetenzen im Umgang mit KI-gestützten Systemen.

Wie rasant die Entwicklung der KI-Technologie vorangeht, zeigt sich bereits daran, dass eine mittlerweile überarbeitete Version von GPT-4.0 (Stand: März 2024) kaum noch oben genannte „einfache“ Rechenfehler macht. Sobald ChatGPT erkennt, dass eine Rechenaufgabe gelöst werden muss, wird nun automatisch ein Plugin gestartet, welches ein Script mit der Programmiersprache Python erzeugt und die Rechenaufgaben häufig fehlerfrei löst.

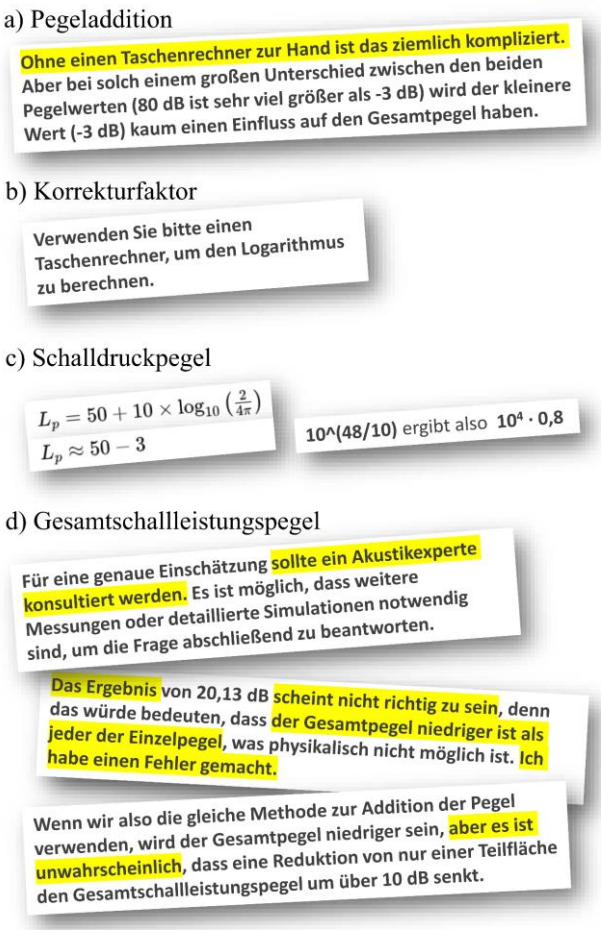


**Abbildung 5:** Anekdotische Schnipsel aus allen gespeicherten Antworten von ChatGPT zu Fragen aus den vier Kategorien.

## Literatur

[1] OpenAI: GPT-4 Technical Report, 2023 https://doi.org/10.48550/arXiv.2303.08774

[2] ZEIT Online „Wie gut können Sie rechnen?“, URL: `https://www.zeit.de/2023/45/mathekenntnisse-deutschland-matheaufgaben-test`

[3] Wei, J. et al.: Emergent Abilities of Large Language Models. Transactions on Machine Learning Research, 08/2022, https://doi.org/10.48550/arXiv.2206.07682

[4] Sengpiel, E.: Falsche Abnahme vom Schalldruck mit der Entfernung von der Schallquelle, URL: `https://sengpielaudio.com/FalscheAbnahmeDesSchalldrucks.htm`

[5] OpenAI ChatGPT, URL: `https://chat.openai.com/`