Eine Vorstudie zur Eignung von Llama 3-8B für eine Sentimentanalyse

Tu, Ngoc Duyen Tanja
tu@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland
ORCID: 0000-0002-7586-0617

Einleitung

Änderungen im deutschen Sprachgebrauch lösen bei vielen Menschen große Emotionen aus, siehe u. a. hitzige Diskussionen zu Reformbestrebungen zu Groß- und Kleinschreibung (Küppers, 1984) oder Proteste gegen die Rechtschreibreform 1998 (Johnson, 2000). Nachdem die deutsche Bundesregierung 2018 das Gesetz verabschiedet hat, im Geburtenregister auch „divers“ eintragen zu können oder keine Geschlechtsangabe zu machen, wurde der Rat für deutsche Rechtschreibung (fortfolgend: Rechtschreibrat) vielfach zum Thema geschlechtergerechte Schreibung angefragt. Im Juli 2024 wurde ein Passus dazu in das Amtliche Regelwerk des Rechtschreibrats aufgenommen. Allerdings wird darin keine klare Empfehlung gegeben, sondern festgehalten, dass der Bereich weiterhin beobachtet wird (Geschäftsstelle des Rats für deutsche Rechtschreibung, 2024, 153–154). Aufgrund dessen ist künftig weiterhin mit Anfragen zur geschlechtergerechten Schreibung zu rechnen. Dabei ist die emotionale Haltung der Anfragenden von großem Interesse. Wie sind die Personen zur geschlechtergerechten Schreibung eingestellt? Welche Emotionen zeigen sich in den Anfragen? Welche Argumente bringen die Positiv- bzw. Negativeingestellten? Um diese Fragen zu beantworten, kommt zunächst eine Sentimentanalyse in Betracht. Mit dieser kann jede Anfrage satzweise mit den Werten „positiv“, „negativ“ oder „neutral“ klassifiziert werden. Im Anschluss können die Sentimentwert-annotierten Daten analysiert werden und es kann in Vorbereitung für eine Emotionsanalyse identifiziert werden, welche Emotionen darin zu finden sind.

In diesem Beitrag wird eine Vorstudie präsentiert, in der getestet wird, ob sich die Generative Künstliche Intelligenz (GenKI) Llama-3-8B Q4_0 instruction-tuned (fortfolgend: Llama-3; Meta, 2024) für eine Sentimentanalyse eignet und für zukünftige Anfragen eingesetzt werden kann. Es wird eine GenKI getestet, da diese eine All-in-one-Solution bietet: Sowohl die Satzsegmentierung als auch die Sentimentklassifikation können mit ihr durchgeführt werden, sowie ggf. weitere Klassifikationen. Kenntnisse in Programmierung müssen dabei nicht vorhanden sein, da man die GenKI mit einer natürlichsprachlichen Eingabe (Prompt) bedient.

Forschungsstand

Studien zu Sentimentanalyse mit einer GenKI wurden bereits durchgeführt: Krugmann und Hartmann (2024) sowie Zhang et al. (2024) zeigen für englische Datensätze, dass die GenKI GPT-4 bzw. Flan-UL2 und GPT-3.5-turbo ähnlich gut oder besser abschneiden als Sprachmodelle, die mit Sentimentwert-annotierten Daten trainiert sind. Zudem stellen Krugmann und Hartmann (2024) fest, dass Llama 2 die besten Erklärungen für seine gewählten Sentimentwerte gibt.

De Araujo et al. (2024) zeigen für brasilianisch portugiesische Datensätze, dass GPT-3.5 bei einer Sentimentanalyse genauso gut oder besser abschneidet wie das bisherige State-of-the-Art-Modell GBT. Zur gleichen Erkenntnis kommen auch Al-Thubaity et al. (2023) für GPT-4 und Bard AI in ihrer Studie mit arabischsprachigen Daten.

Zu anderen Ergebnissen kommen Mæhlum et al. (2024) für ChatNort4 und NorMistral sowie Rønningstad et al. (2024) für GPT-4, die mit norwegischen Datensätzen arbeiten. Sie zeigen, dass die GenKI schlechter annotieren als Menschen. Für einen englischen Datensatz zeigen Møller et al. (2024), dass GPT-4 und Llama 2-70-B-Chat nicht so gut abschneiden wie Sprachmodelle, die mit Sentimentwert-annotierten Daten trainiert sind. Gleiches zeigen Pfister und Hotho (2024) für einen deutschen Datensatz mit german-gpt2, einer auf ca. 2 Mrd. Token deutschen Texten trainierten GenKI.

Bisher finden sich keine Untersuchungen zu Sentimentanalysen auf deutschen Texten mit GenKI, die auf sehr großen Datenmengen trainiert sind. Dafür soll die vorliegende Untersuchung einen ersten Anhaltspunkt geben.

Studienkonzeption

Datengrundlage

Als Datengrundlage dienen 30 von 146 zufällig gezogenen, manuell anonymisierten E-Mail-Anfragen von 2019-02/2024 an den Rechtschreibrat, die insgesamt aus 233 Sätzen bestehen. Im Durchschnitt besteht eine Anfrage aus 134 Token und 8 Sätzen. Alleinstehende Grußformeln wie *Sehr geehrte* oder *Viele Grüße* wurden entfernt, da sie als Floskeln für die Sentimentanalyse irrelevant sind.

Die Datengrundlage liegt aus datenschutzrechtlichen Gründen nicht öffentlich vor. Dies ist für die Studie vorteilhaft, da die Performance von Llama-3 getestet wird und es somit nicht zu einer Datenkontamination kommen kann, d. h. es kann ausgeschlossen werden, dass der Datensatz bereits annotiert in den Trainingsdaten der GenKI vorliegt.

Vorgehensweise

GenKI

Für die Sentimentanalyse wird die, nach seinem Entwickler Meta, leistungsstärkste Open-Source-GenKI Llama 3 über die Python-Bibliothek ollama genutzt. In Benchmark-Tasks zeigt Llama 3 gegenüber den Open-Source-GenKI Gemma-7B-It und Mistral-7B-Instruct eine höhere Performance (Meta, 2024).

Ein großer Vorteil von Llama 3 gegenüber Close-Source-GenKI ist, dass man es lokal auf dem eigenen Rechner ausführen kann, womit man die Kontrolle über die Datengrundlage behält.

Prompt

Studien belegen, dass die Wahl des Prompts ausschlaggebend für die Qualität der Antwort ist (Battle und Gollapudi, 2024; Leidinger et al., 2023; Schulhoff et al., 2024; Zhang et al., 2024). Bsharat et al. (2023) zeigen, dass die Performance der Modelle Llama-1/-2 sowie die von GPT-3.5/-4 bei Befolgung bestimmter Regeln zur Formulierung des Prompts steigt. Entsprechend wurden folgende, zum Task passende, Regeln aus Bsharat et al (2023) angewendet, um den Prompt zu formulieren: i) Unterlassung von Höflichkeitsfloskeln; ii) Nutzung von Direktiven; iii) Kenntlichmachung zusammenhängender Abschnitte im Prompt; iv) Nutzung der Phrasen „Your task is“ und „You must“ und v) mehrfache Wiederholung bestimmter Wörter oder Phrasen. Semantisch orientiert sich der Prompt an genutzte Prompts für Sentimentanalysen mit GenKI aus anderen Studien (De Araujo et al., 2024; Krugmann und Hartmann, 2024; Mæhlum et al., 2024; Zhang et al., 2024; Al-Thubaity et al., 2023):

> ###Instruction### Your task is to do a sentiment analysis (positive, negative, neutral) for each sentence in the following text. You must assign only one sentiment value (positive, negative or neutral). You must give your answer in the format “Sentence” -- “Sentiment value” -- “Justification”. ###Text### + *question*

Pro Prompt wird eine Anfrage aus der Datengrundlage präsentiert, indem sie an den Prompt am Ende konkateniert wird (+ question).

Die Instruktion wird auf Englisch gegeben, da die Trainingsdaten von Llama-3 zu 95 % aus englischen Texten besteht (Meta, 2024) und die Antwortqualität deutlich schlechter bei nicht-englischsprachigen Prompts ausfallen kann (Schulhoff et al., 2024). Die Anfrage wird auf Deutsch übergeben, um dem Problem zu entgehen, dass sprachliche Nuancen bei einer Übersetzung verloren gehen. Testweise wurde der Prompt vollständig auf Deutsch formuliert, was zu schlechteren Ergebnissen bei der Satzsegmentierung geführt hat.

Damit Llama-3 Kontextinformationen erhält, die bei der Sentimentklassifikation relevant sein können, wird ihm die gesamte Anfrage im Prompt übergeben und es führt zunächst eine Satzsegmentierung durch. Danach soll die GenKI die Sentimentklassifikation vornehmen und anschließend eine Begründung für den vergebenen Sentimentwert geben. Diese Reihenfolge der Instruktionen wurde gewählt, da der GenKI bei der Textgenerierung nur die bereits generierten Tokens bekannt sind und somit konsistente Antworten erwartet werden. Somit soll also bewirkt werden, dass der vergebene Sentimentwert und die Begründung der GenKI in einem logischen Zusammenhang zueinanderstehen.

Eine Systemmessage wird nicht gesetzt. Bei der Temperatur, die angibt, wie kreativ die GenKI antwortet, wobei 1 für sehr kreativ und 0 für deterministisch steht, wird der default-Wert von 0,6 beibehalten.

In 5 Fällen vergibt Llama-3, abweichend von der Instruktion, 2 Sentimentwerte, in diesen Fällen wurden von mir automatisch der erste Sentimentwert gesetzt.

Baselines

Als Baseline dient das Sprachmodell german-sentiment-bert (Guhr et al., 2020), das auf deutschen Sentimentwert-annotierten Daten weitertrainiert ist. Dieses Modell wurde gewählt, da die Trainingsdaten aus nicht-redigierten Texten wie Facebook-Posts bestehen und die vorliegende Datengrundlage ebenfalls nicht-redigierte Texte enthält. Zusätzlich wurde zur Einordnung der Ergebnisse eine random-Baseline berechnet, die den Sätzen aus der Datengrundlage zufällig Sentimentwerte zuordnet.

Manuelle Annotation

Für die manuelle Annotation wurden drei Aufgaben gestellt[1]:

(1) Drei Linguist:innen (M0, M1, M2) bekommen die Anfragen, die von Llama-3 Satz-segmentiert wurden. Es ist ersichtlich, welche Sätze zu einer Anfrage gehören. Ihre Aufgabe ist es, die Sätze mit einem Sentimentwert zu annotieren, dabei können sie Kontextinformationen einbeziehen. Die Annotationsbedingungen sind somit die gleichen wie die für Llama-3.

(2) In einem 2. Schritt werden den drei Linguist:innen jeweils die Annotationen von Llama-3 präsentiert, die von ihren eigenen abweichen. Sie entscheiden, ob sie die Begründung von Llama-3 für den gewählten Sentimentwert überzeugend finden.

(3) Drei weitere Linguist:innen (M3, M4, M5) bekommen den Datensatz, der wie in (1) beschrieben aufbereitet ist, die Annotationen und die Begründungen von Llama-3. Ihre Aufgabe ist es zu entscheiden, ob sie der Annotation von Llama-3 zustimmen. Wenn sie nicht zustimmen, vergeben sie einen anderen Sentimentwert.

Ergebnisse

Inter-Annotator-Agreement

Basierend auf dem Cohens Kappa κ (Cohen, 1960) wurde das Inter-Annotator-Agreement bestimmt. Ergebnisse von κ liegen zwischen -1 (potenziell systematische Nichtübereinstimmung) und 1 (perfekte Übereinstimmung).

Zunächst wird κ für die Annotationsaufgabe (1) berechnet (vgl. Tabelle 1).

Tabelle 1: Das IAA zwischen den menschlichen Annotierenden (M0, M1, M2) untereinander, Llama-3 sowie den Baselines BERT und random.

Annotierender	M0	M1	M2
M0	X	0,56	0,35
M1	0,56	X	0,53
M2	0,35	0,53	X
Llama-3	0,46	0,35	0,25
BERT	0,10	0,23	0,18
random	0,03	-0,01	-0,03

Aus Tabelle 1 geht hervor, dass das Inter-Annotator-Agreement zwischen den menschlichen Annotierenden und Llama-3 höher ist als zwischen den Baselines. Des Weiteren stimmen sowohl die Menschen untereinander als auch mit Llama-3 jeweils nur mittelmäßig bis moderat überein (Viera und Garrett, 2005, 362).

Als Nächstes wird κ für die Annotationsaufgabe (3) berechnet (vgl. Tabelle 2).

Tabelle 2: Das IAA zwischen den menschlichen Annotierenden (M3, M4, M5) untereinander, Llama-3 sowie den beiden Baselines BERT und random.

Annotierender	M3	M4	M5
M3	X	0,59	0,49
M4	0,59	X	0,42
M5	0,49	0,42	X
Llama-3	0,69	0,50	0,37
BERT	0,13	0,11	0,05
random	0,03	0	0,06

Aus Tabelle 2 geht hervor, dass das Inter-Annotator-Agreement zwischen den menschlichen Annotierenden und Llama-3 ebenfalls höher ist als zwischen den Baselines. Vergleicht man die Werte aus Tabelle 1 mit denen aus Tabelle 2 kann man sehen, dass die κ-Werte bei Aufgabe (3) höher sind. Jedoch variieren sie zwischen den Menschen und Llama-3 stark: M3 weist eine substantielle Übereinstimmung mit Llama-3 auf, M5 nur eine mittelmäßige. M4 liegt dazwischen mit einer moderaten Übereinstimmung.

Insgesamt kann aus den κ-Werten folgendes abgeleitet werden:

- Die κ-Werte in Tabelle 1 für Annotationsaufgabe (1) weisen darauf hin, dass die Annotationsrichtlinien nicht präzise genug sind, weshalb das Inter-Annotator-Agreement zwischen den menschlichen Annotierenden nicht so hoch ist. M0 und M1 haben zurückgemeldet, dass sie die Annotationsaufgabe als schwierig empfanden, da das Thema selbst bereits emotionsbeladen ist. Somit bekämen die meisten Sätze ihres Empfindens nach meist einen impliziten negativen Ton. Die Aussage wird durch die Verteilung der vergebenen Sentimentwerte je Annotierende belegt (vgl. Tabelle 3): M0 und M1 klassifizieren im Vergleich zu Llama-3 deutlich mehr Sätze als „negativ".

- Die stark variierenden κ-Werte in Tabelle 2 für Annotationsaufgabe (3) weisen darauf hin, dass die Annotationen von Llama-3 je nach subjektivem Empfinden überzeugend sein können. Aus Tabelle 3 geht hervor, dass M4 und M5, die ein niedriges Inter-Annotator-Agreement mit Llama-3 haben, deutlich mehr Sätze als „negativ" annotiert haben als Llama-3. Zusätzlich hat M5 deutlich weniger Sätze als „positiv" annotiert.

- Darüber hinaus ist es nicht überraschend, dass der Großteil der Sätze als neutral annotiert wurde, da Personen die eine Anfrage an den Rechtschreibrat stellen, oftmals eine Antwort auf ein sprachliches – neutral dargebrachtes – Problem suchen, z. B. *Wie ist denn die bestehende Regelung für Behörden* […].

Zusammenfassend ist zu beobachten, dass, auf Grund der Thematik der Datengrundlage, viele subjektive Empfindungen in die Sentimentklassifikation einfließen, weshalb auch das Inter-Annotator-Agreement zwischen den menschlichen Annotierenden nicht besonders hoch ist. Folglich stellt sich als Nächstes die Frage, ob die vergebenen Sentimentwerte von Llama-3 ebenfalls plausibel sind. Um diese Frage zu beantworten, werden die Ergebnisse von Annotationsaufgabe (2) betrachtet.

Tabelle 3: Die Anteile der vergebenen Sentimentwerte je Annotierender.

Sentim-entwert	M0	M1	M2	M3	M4	M5	Llama-3	BERT	random
positiv	32	16	12	58	54	21	68	15	80
negativ	66	66	34	49	73	76	38	21	68
neutral	135	151	187	126	106	136	127	197	85

Qualitative Analyse der Begründungen von Llama-3

Aus Tabelle 3 geht hervor, dass Llama-3 im Gegensatz zu M0, M1 und M2 deutlich mehr Sätze als „positiv" klassifiziert hat. Betrachtet man die Sätze, die Llama-3 abweichend von den drei Annotierenden als „positiv" annotiert hat, kann man folgende Beobachtungen machen:

- In 28 Sätzen finden sich positive Emotionen wie *freuen*, das auf Grund seines Vorkommens in einer Floskel: *Ich würde mich über baldige Antwort sehr freuen.* von den menschlichen Annotierenden mehrheitlich als „neutral" annotiert wurde. Die Begründung der GenKI: „The speaker's enthusiasm for a prompt response indicates a positive sentiment." überzeugt die Annotierenden von einer positiven Annotation.

- In 10 Sätzen werden positive Veränderungen beschrieben. Die Annotierenden klassifizieren diese Sätze mehrheitlich als „neutral", finden die Begründungen von Llama-3 jedoch überzeugend. Ein Beispiel hierfür lautet: *Eltern könnten ihren Kindern wieder bei den Hausaufgaben helfen,* […], wobei die Begründung von Llama-3 wie folgt lautet: „The speaker suggests that the proposed changes would allow parents to help their children more effectively, which is a positive sentiment."

- Es ist auffällig, dass Llama-3 Sätze als „positiv“ klassifiziert, wenn sie eine ‚kämpferische‘ Handlung enthalten. Darunter fallen 10 Sätze, wie beispielsweise *Ich* […] *kämpfe seit langem gegen diese merkwürdigen Auswüchse,* […] . Llama-3 begründet die Annotation folgendermaßen: „The speaker expresses determination to fight against these “weird” developments […].)” Die Begründungen bei diesen Sätzen finden die Annotierenden nicht überzeugend und klassifizieren sie mehrheitlich als „negativ“.

- Die Begründungen von 3 Sätzen, die von Llama-3 als „positiv“ klassifiziert wurden, finden die Annotierenden nicht überzeugend. Die Problematik bei diesen Sätzen liegt darin, dass man ihnen, je nachdem, welchen Aspekt man für die Sentimentklassifikation einbezieht, einen anderen Sentimentwert zuordnen würde. Einer dieser Sätze lautet: *Das wäre eine klar strukturierte und verständliche Lösung ohne [Verunstaltung] der Deutschen Sprache.* Llama-3 begründet die „positive”-Klassifikation mit: „The use of words like “klar” (clear) and “verständlich” (understandable) suggests a positive sentiment towards the proposed solution.” Die Annotierenden klassifizieren diesen Satz mehrheitlich als “neutral”, da ein Lösungsvorschlag für geschlechtergerechte Schreibung präsentiert wird. Auf Grund der Verwendung des Wortes *Verunstaltung* wäre auch eine Klassifikation als „negativ“ denkbar.

Bei der Analyse weiterer Sätze, bei denen sich die menschlichen Annotierenden und Llama-3 uneinig sind, wird sichtbar, dass in diesen sehr viele Aspekte vorkommen, die mit Sentimentwerten klassifiziert werden könnten. Darunter fallen beispielsweise die Sprechweise, die Haltung des Anfragenden oder Emotionen, Situationen und Handlungen, die im Satz beschrieben werden. Dieser Umstand erschwert die Sentimentanalyse bei dieser Studie.

Fazit

In diesem Beitrag wurde gezeigt, dass sich Llama-3 mit den in dieser Vorstudie genutzten Annotationsrichtlinien nicht dazu eignet, eine Sentimentanalyse durchzuführen. Die Analyse der Begründungen von Llama-3 hat gezeigt, dass die Annotationsrichtlinien verfeinert werden müssen: Es muss spezifiziert werden, welche Aspekte in die Sentimentklassifikation einfließen sollen. Somit kann Llama-3 in den Digitalen Geisteswissenschaften bei der Erstellung sowie der Verfeinerung von Annotationsrichtlinien genutzt werden. Damit kann zunächst die Arbeitskraft von menschlichen Annotierenden gespart werden. Durch die GenKI können Aspekte aufgedeckt werden, die eine Einzelperson (auf Grund subjektiver Empfindungen) nicht präsent hat.

In einem nächsten Schritt sollte eine Folgestudie durchgeführt werden, in dem die Annotierenden und Llama-3 entsprechend der verfeinerten Annotationsrichtlinien instruiert werden. Es bleibt abzusehen, ob sich das IAA dadurch erhöhen wird und Llama-3 für den Einsatz einer Sentimentanalyse in Frage kommt.

Fußnoten

1. Vielen Dank an die Annotierenden: Jennifer Behr, Beate Brechtel, Annelen Brunner, Christian Lang, Niklas Reinken und Roman Schneider.

Bibliographie

Al-Thubaity, Abdulmohsen, Sakhar Alkhereyf, Hanan Murayshid, Nouf Alshalawi, Maha Omirah, Raghad Alateeq, Rawabi Almutairi, Razan Alsuwailem, Manal Alhassoun und Imaan Alkhanen. 2023. "Evaluating ChatGPT and Bard AI on Arabic Sentiment Analysis." In *Proceedings of ArabicNLP 2023*, 335–349.

de Araujo, Gladson, Tiago de Melo und Carlos Maurício S. Figueiredo. 2024. "Is ChatGPT an effective solver of sentiment analysis tasks in Portuguese? A Preliminary Study." In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, 13–21.

Battle, Rick und Teja Gollapudi. 2024. "The Unreasonable Effectiveness of Eccentric Automatic Prompts." arXiv. doi:10.48550/ARXIV.2402.10949, https://arxiv.org/abs/2402.10949 (zugegriffen: 2. Juli 2024).

Bsharat, Sondos Mahmoud, Aidar Myrzakhan und Zhiqiang Shen. 2023. "Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4." arXiv. doi:10.48550/ARXIV.2312.16171, https://arxiv.org/abs/2312.16171 (zugegriffen: 2. Juli 2024).

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20, Nr. 1: 37–46.

Geschäftsstelle des Rats für deutsche Rechtschreibung, Hrsg. 2024. *Amtliches Regelwerk der deutschen Rechtschreibung. Regeln und Wörterverzeichnis*. Mannheim: IDS-Verlag.

Guhr, Oliver, Anne-Kathrin Schumann, Frank Bahrmann und Hans Joachim Böhe. 2020. "Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems." In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1627–1632.

Johnson, Sally. 2000. "The Cultural Politics of the 1998 Reform of German Orthography." *German Life and Letters* 53, Nr. 1: 106–125.

Krugmann, Jan Ole und Jochen Hartmann. 2024. "Sentiment Analysis in the Age of Generative AI." *Customer Needs and Solutions* 11: 3.

Küppers, Hans-Georg. 1984. "Orthographiereform und Öffentlichkeit: zur Entwicklung und Diskussion der Rechtschreibreformbemühungen zwischen 1876 und 1982." Sprache der Gegenwart 61. Düsseldorf: Schwann.

Leidinger, Alina, Robert Van Rooij und Ekaterina Shutova. 2023. "The language of prompting: What linguistic properties make a prompt successful?" In:

Findings of the Association for Computational Linguistics: EMNLP 2023, 9210–9232.

Mæhlum, Petter, David Samuel, Rebecka Maria Norman, Elma Jelin, Øyvind Andresen Bjertnæs, Lilja Øvrelid und Erik Velldal. 2024. "It's Difficult to Be Neutral – Human and LLM-based Sentiment Annotation of Patient Comments." In: *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, 8–19.

Meta. 2024. Build the future of AI with Meta Llama 3. https://llama.meta.com/llama3/.

Pfister, Jan und Andreas Hotho. 2024. "SuperGLEBer: German Language Understanding Evaluation Benchmark." In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7904–7923.

Rønningstad, Egil, Erik Velldal und Lilja Øvrelid. 2024. "A GPT among Annotators: LLM-based Entity-Level Sentiment Annotation." In: *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, 133–139.

Schulhoff, Sander, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, u. a. 2024. "The Prompt Report: A Systematic Survey of Prompting Techniques." arXiv. doi:10.48550/ARXIV.2406.06608, https://arxiv.org/abs/2406.06608 (zugegriffen: 2. Juli 2024).

Viera, Anthony J. und Joanne M. Garrett. 2005. "Understanding Interobserver Agreement: The Kappa Statistic." *Family Medicine* 37, Nr. 5: 360–363.

Zhang, Wenxuan, Yue Deng, Bing Liu, Sinno Pan und Lidong Bing. 2024. "Sentiment Analysis in the Era of Large Language Models: A Reality Check." In: *Findings of the Association for Computational Linguistics*, 3881–3906.