Empirische Evaluation des Verhaltens von LLMs auf Basis sprachphilosophischer Theorien: Methode und Pilotannotationen

Pichler, Axel
apichler@ts.uni-stuttgart.de
Universität Stuttgart, Deutschland

Gerstorfer, Dominik
dominik.gerstorfer@tu-darmstadt.de
Technische Universität, Darmstadt, Deutschland

Kuhn, Jonas
jonas.kuhn@ims.uni-stuttgart.de
Universität Stuttgart, Deutschland

Pagel, Janis
janis.pagel@uni-koeln.de
Universität zu Köln, Deutschland
ORCID: 0000-0003-4370-1483

Einleitung

Die Fähigkeit großer Sprachmodelle (LLMs), Bedeutung und Verstehen zu simulieren, hat in den letzten Jahren zu einer regen Debatte darüber geführt, inwiefern LLMs tatsächlich bedeutungsvolle Sprache generieren und diese verstehen. Bedeutungs- und Verstehensbegriff sind zugleich zentrale Konzepte der Kultur- und Geisteswissenschaften. Dementsprechend beteiligen sich auch viele traditionelle und digitale Geisteswissenschaftler:innen an der laufenden Debatte. So wurde 2023 auf einem Panel der DHd-Konferenz ein Zugang zu dieser Debatte präsentiert, der das Ziel hatte, durch theoretische Impulse eine präzisere Beschreibungssprache für die Digital Humanities (DH) zu schaffen, die dabei helfen sollte, die Unterschiede zwischen den Bedeutungsprozessen von Maschinen und Menschen herauszuarbeiten, um so die Anwendung von Künstlicher Intelligenz kritisch zu hinterfragen (Gengnagel et al. 2024). Während das Panel derartig eine geisteswissenschaftliche Kernkompetenz – die Begriffsanalyse – reaktivierte, um ordnend in die Diskussion einzugreifen, inwiefern LLMs bedeutungsvolle Sprache generieren und diese verstehen, folgte es zugleich der auch in der Natural-Language-Processing-Community weitverbreiteten Tendenz, besagte Fragen primär theoretisch zu verhandeln.

Wir wollen dieser theoretischen Debatte keine weitere theoretische Position hinzufügen, sondern im Folgenden einen Vorschlag machen, wie ergänzend überprüft werden kann, inwiefern das Verhalten eines LLMs existierenden Sprachtheorien entspricht. Dabei knüpfen wir an die NLP-Tradition des ›behavioral testings‹ an, die sich mit der Prüfung verschiedener Fähigkeiten eines Systems durch Validierung des Eingabe-Ausgabe-Verhaltens ohne Kenntnis der internen Struktur befasst (Beizer 1995). Hierbei verfolgen wir einen theoriegeleiteten Top-Down-Ansatz, der von einer gegebenen Sprachtheorie ausgeht und diese so modelliert, dass ihre zentralen Begriffe in einer Form operationalisiert werden können (Krautter/Pichler/Reiter 2023; Gerstorfer/Gius 2023), welche die Erstellung eines Testdatensatzes erlaubt, der die zentralen sprachtheoretischen Annahmen der Referenztheorie in einem angemessenen Grad repräsentiert. Im Falle von Bedeutungstheorien sollte ein derartiger Testdatensatz das Output eines kompetenten Sprechers auf eine Art und Weise abbilden, wie es von der untersuchten Sprachtheorie impliziert wird. Mit Hilfe eines solchen Testdatensatzes könnte dann der Grad bestimmt werden, in dem das Sprachverhalten eines LLMs dem einer Sprachtheorie entspricht. Ein derartiges Wissen über das Verhalten von LLMs ist insbesondere für jene Zweige der DH relevant, deren Theorien und Analysen auf bestimmten sprachphilosophischen Vorannahmen aufbauen. Sie könnten dann für ihre Analysen jene LLMs verwenden, die diesen entsprechen.

Wir werden daher im Folgenden eine Methode präsentieren, die eine derartige Evaluation erlaubt, sie anhand eines Beispiels – der wahrheitskonditionalen Semantik von Donald Davidson – vorführen, und die Resultate der ersten Pilotannotationen sowie erster Experimente mit LLMs präsentieren.

An diesem Punkt sei darauf hingewiesen, dass die Auswahl von Davidsons Sprachtheorie nicht daher rührt, dass wir glauben, dass sie in höherem Grad als alternative Sprachtheorien dem Textgenerierungsverhalten von LLMs entspricht, sondern daher, dass Ausgangsszene und Kernkonzept seiner Sprachphilosophie sehr treffend die Situation beschreiben, mit der Benutzer von großen Sprachmodellen konfrontiert sind: die radikale Interpretation. Bei dieser steht der radikale Interpret einer Sprecher:in einer ihm unbekannten Sprache gegenüber und versucht auf Basis einer spezifischen Form des *principles of charity* sowie unter Berücksichtigung der gegebenen Situation, die durch diese Situation konditionierten Äußerungen seines Gegenübers zu interpretieren und zu verstehen. Überträgt man diese Situation auf die Interaktion mit LLMs, führt dies dazu, dass die von einem LLM generierten sprachlichen Äußerungen wie Äußerungen einer fremden Sprache rezipiert werden, die radikal interpretiert werden müssen, um verstanden werden zu können.

Abgesehen von diesen Parallelen ist die im Folgenden vorgestellte Methode zur Entwicklung eines Testdatensatzes zur Überprüfung, inwiefern das Verhalten eines LLMs

den Erwartungen einer Sprachtheorie bezüglich des Verhaltens eines kompetenten Sprechers entspricht, sprachtheorie-agnostisch. Mit ihrer Hilfe können auch alternative Sprachtheorien getestet werden, was in Anbetracht der Vielfalt des Theorieangebots in Sprachphilosophie und Semantik sowie von deren zentraler Rolle zum Beispiel in den Interpretationstheorien der Literaturwissenschaft ein Forschungsdesiderat darstellt. Längerfristig streben wir an, weitere sprachphilosophische Theorien derartig zu überprüfen.

Methode

Die Generierung eines Testdatensatzes zur Überprüfung, inwiefern das Verhalten eines LLMs den Erwartungen einer Sprachtheorie entspricht, erfolgt in drei Schritten: In einem ersten Schritt sind die zentralen Annahmen der zu testenden Sprachtheorie rational zu rekonstruieren. Ziel dieser Rekonstruktion ist, zweitens, die Rückführung besagter Sprachtheorie auf eine oder mehrere sprachtheoretische Hypothesen, die im Folgenden getestet werden. Dafür sind, drittens, die zentralen Begriffe dieser Hypothesen so zu operationalisieren, dass mit ihrer Hilfe ein Testdatensatz erzeugt werden kann.

Im Zentrum von Davidsons Bedeutungstheorie steht die These, »that a theory of truth, modified to apply to a natural language, can be used as a theory of interpretation« (Davidson 2006, S. 189). Davidson kommt einem Ansatz wie dem hier präsentierten, der auf eine operationalisierbare Rekonstruktion einer Theorie abzielt, nun insofern entgegen, als dass Davidson selbst mit dem Konzept der radikalen Interpretation bereits eine Operationalisierung der Kernelemente seiner Sprachtheorie vorgelegt hat, die nur in Hinblick auf jene Voraussetzungen zu adaptieren ist, die LLMs im Unterschied zu kompetenten menschlichen Sprechern nicht erfüllen. Davidson schreibt: »A theory of meaning (in my mildly perverse sense) **is an empirical theory**, and its ambition is to account for the workings of a natural language. Like any theory, it may be tested by comparing some of its consequences with the facts. In the present case this is easy, for the theory has been characterized as issuing in an infinite flood of sentences each giving the truth conditions of a sentence; we only need to ask, in sample cases, whether what the theory avers to be the truth conditions for a sentence really are. A typical test case might involve deciding whether the sentence ‘Snow is white’ is true if and only if snow is white.« (Davidson 2006, S. 161, Hervorhebung von den Autoren)

In Hinblick auf die zu konstituierende Leithypothese heißt das, dass einem Sprachmodell eine Vielzahl von Gelegenheitssätzen in Bezug auf eine bestimmte Situation vorzulegen ist, um dann zu überprüfen, inwiefern das Model diese Sätze für wahr hält. Die absurd anmutende Formulierung des soeben artikulierten verweist bereits auf jene Elemente der Davidson’schen Theorie, die zu adaptieren sind, wenn man sie auf LLMs anwenden möchte. Dazu zählen insbesondere, dass 1.) große Sprachmodelle keine Agenten sind, 2.) keine Überzeugungen besitzen und dementsprechend 3.) auch keine Propositionen für-wahr-halten können. LLMs können jedoch das entsprechende Verhalten eines kompetenten Sprechers simulieren. Zudem entspricht die ›Kommunikationssituation‹ zwischen einem LLM und einem Menschen nicht derjenigen der radikalen Interpretation: Weder besitzt eine solche einen realweltlichen Situations- und Bezugsrahmen, auf Basis dessen in Übereinstimmung mit dem Gesamtverhalten eines Sprechers einer fremden Sprache bestimmt werden kann, ob dieser einen Satz zu einem bestimmten Zeitpunkt an einem bestimmten Ort für wahr hält oder nicht, noch handelt es sich dabei um eine kausale Relation zwischen Bezugsrahmen und Verhalten besagten Sprechers (vgl. Davidson 2001). Das im Folgenden entwickelte Testset überprüft dementsprechend nur, inwiefern sich diese Simulation zum Kommunikationsverhalten eines kompetenten Sprechers im Sinne Davidsons verhält bzw. zu welchem Grade es diesem entspricht. Ausgangspunkt bei der Entwicklung des Testsets ist dabei folgende Leit-Hypothese: (H) Ein Sprachmodell verwendet eine Sprache wie ein kompetenter Sprecher im Sinne Davidsons, gdw. es die selben zum Zeitpunkt Z geäußerten Sätze im Verhältnis zum sprachlichen Kontext K als wahr bestimmt.

Testdatensatzerstellung

Die Testdatenerzeugung erfolgt dementsprechend in Bezug auf einen sprachlichen Kontext K, der von jedem beliebigen Text gefüllt werden kann, der eine Situation beschreibt, die im Hinblick auf situative Aussagen auf ihren Wahrheitswert hin überprüft werden kann. Dieser Fokus auf wahrheitskonditionale situative Aussagesätze ist der Orientierung an den theoretischen Grundannahmen der Davidson’schen Theorie geschuldet. Im Falle alternativer sprachphilosophischer Theorien können andere Frage-Antwort-Typen relevant sein.

Für unsere Pilotannotation haben wir auf die Texteröffnung von Franz Kafkas Erzählung *Das Urteil* zurückgegriffen, die laut Michael Scheffel als »fiktionale, illusionistische, autor- und erzählerverleugnende, aliozentrische Autorerzählung in dritter Person« mit Dietrich Weber als Standardtyp des Erzählens verstanden werden kann (Scheffel 2002, S. 61). Diese Auswahl ist eine Folge unserer leitenden Forschungsinteressen. Eine extensivere Testung von Sprachmodellen, die umfassendere Geltungsansprüche stellte als die hier vorgelegte Pilotstudie, müsste mit einer Vielzahl derartiger Kontexte arbeiten und dabei deren Auswahl ebenfalls auf Basis der untersuchten Sprachtheorie plausibilisieren. Davidson gliedert nun den Prozess für die Erarbeitung einer Wahrheits- bzw. Bedeutungstheorie für eine unbekannte Sprache, auf welche die Quantorenlogik erster Stufe projiziert wird, in drei Schritte: 1. Identifizierung (unter anderem) aller Prädikate und Singular-Phrasen, 2. Erstellung von Sätzen mit Wahrheitsbedingungen bezüglich eines gegebenen Kontextes und 3. Sätze, die nicht entscheidbar sind (vgl. Davidson 2006, S. 193).

Wir haben uns für unsere Experimente auf die ersten beiden der drei genannten Schritte konzentriert und dementsprechend in einem ersten Schritt aus den beiden Textstellen das Basisvokabular extrahiert, indem wir mithilfe von spaCy[1] sämtliche transitiven und intransitiven Verben sowie die Nominalphrasen aus dem Text zogen. In einem zweiten Schritt haben wir auf Basis dieses Vokabulars semiautomatisch 200 situative Aussagesätze generiert und zwar 175 mithilfe von GPT-4o[2] und einem Prompt, der das Modell dazu aufforderte, 25 Beispiele von sieben vorgegebenen Satzformen zu generieren. In einem zweiten Schritt wurde das Modell angewiesen, die solcherart generierten Sätze in eine grammatisch korrekte Form zu überführen.[3] Ergänzt wurde dieser Bestand um 25 manuell erzeugte Sätze, die in Bezug auf die im Text beschriebene Situation wahr sein sollten, da wir davon ausgehen, dass eine große Mehrheit der erzeugten Sätze bezüglich des Kontextes K falsch sein werden.

Pilotstudie und LLM-Experimente

Wir führen eine Pilotstudie zur Annotation der Testdaten durch, um die Durchführbarkeit des Vorhabens zu demonstrieren.

Die Testdaten zu Kafkas Urteil wurden von drei Annotatoren (drei der Autoren; die 25 manuellen Sätze wurden vom ersten Autor erstellt, der selber nicht annotiert hat) annotiert, und zwar bezüglich des Wahrheitswertes des Satzes im Hinblick auf den gegebenen Kontext als auch bezüglich der Angabe, ob der Satz extrinsisch oder intrinsisch wahr oder falsch ist.

Die Auswertungen der Annotationen in Form einer Inter-Annotator-Agreement-Studie befinden sich in Tabelle 1 (Agreement bezüglich der Wahrheitswerte) und Tabelle 2 (Agreement bezüglich intrinsisch/extrinsisch). Für die Auswertungen wurden Sätze, die von den Annotatoren als nicht-entscheidbar eingeschätzt wurden, auf die Werte falsch, bzw. intrinsisch gesetzt.

Die Tabellen zeigen das IAA für alle Sätze und für Teilmengen (subsection) der Sätze: (i) ist der Satz ein manuell oder automatisch erstellter Satz, (ii) enthält der Satz einen Junktor oder nicht und (iii) falls der Satz einen Junktor enthält, welchen? Gezeigt wird das resultierende Fleiss' Kappa (Fleiss 1971) als Maß dafür, wie stark die Annotationen übereinstimmen[4] (fleiss.kappa), der p-Wert als statistisches Signifikanzmaß (p.value), die Anzahl an Annotatoren (raters) und die Anzahl an annotierten Sätzen in den Teilmengen (subjects).

Tabelle 1: IAA-Auswertung zu den Wahrheitswerten der Sätze (wahr/falsch).

subsection	fleiss.kappa	p.value	raters	subjects
alle	0.5566922	0.00	3	200
nur automatisch	0.5566151	0.00	3	175
nur manuell	0.3386243	0.003	3	25
nur ohne Junktoren	0.7643636	0.00	3	60
nur mit Junktoren	0.4654545	0.00	3	140
nur Negationen	0.2021277	0.01	3	50
nur Implikationen	0.0865783	0.37	3	35
nur Konjunktionen	0.7452055	0.00	3	31
nur Disjunktionen	0.7473684	0.00	3	24

Tabelle 2: IAA-Auswertung dazu, ob ein Satz extrinsisch oder intrinsisch wahr/falsch ist.

subsection	fleiss.kappa	p.value	raters	subjects
alle	-0.2472789	0.00e+00	3	200
nur automatisch	-0.2466895	0.00e+00	3	175
nur manuell	-0.4722864	1.97e-05	3	25
nur ohne Junktoren	-0.2427966	5.00e-07	3	60
nur mit Junktoren	-0.2500752	0.00e+00	3	140
nur Negationen	-0.2430939	5.30e-06	3	50
nur Implikationen	-0.2333237	1.48e-04	3	35
nur Konjunktionen	-0.3119122	2.44e-05	3	31
nur Disjunktionen	-0.3061224	1.54e-04	3	24

Für die Wahrheitswerte gibt es mit einem Fleiss' Kappa von 0.56 ein moderat gutes Agreement. Am höchsten ist das Agreement für Sätze ohne logische Junktoren und für die Kon- und Disjunktion (Fleiss' Kappa von ca. 0.75). Am wenigsten Übereinstimmungen gibt es für die Implikations-Sätze, wobei das Ergebnis nicht statistisch signifikant ist ($p>0.05$). Für die automatisch erstellten Sätze lässt sich ein etwas höheres Agreement ablesen als für die manuell erstellten. Im Bezug auf extrinsische/intrinsische Wahrheitswertzuschreibungen ist Fleiss' Kappa durchweg negativ, was darauf hindeutet, dass die Annotatoren die Annotationsguidelines unterschiedlich interpretiert haben.

Erste Experimente mit zwei LLMs – OpenAI's GPT-4o und Anthropics Claude 3.5 Sonnet Model[5] – auf denjenigen Daten, bei denen alle Annotatoren übereinstimmen (138 von 200 Sätzen), erzielten einen F1-Score von 85% mit GPT-4o und 77% mit Claude 3.5 Sonnet sowie eine Accuracy von 86% und 79%. Das bedeutet, dass GPT-4o in 86% und Claude 3.5 Sonnet in 79% der Fälle den Gelegenheitssätzen denselben Wahrheitswert zuordnet, wie dies ein kompetenter Sprecher nach Davidson das in Hinblick auf den vorgegebenen Kontext tun sollte.

In summa haben wir gezeigt, dass es möglich ist eine Sprachtheorie so zu modellieren und anschließend in Hinblick auf einen bestimmten Testkontext zu operationalisieren, dass die Resultate als Testdaten für diese Sprachtheorie hinreichen. Im Zuge unserer ersten Experimente haben wir festgestellt, dass eine solche Modellierung und Operationalisierung jedoch zahlreiche Fallstricke besitzt: So führt zum Beispiel eine vollständig automatisierte Testdatengenerierung auf Basis eines gegebenen Vokabulars mehrheitlich zu sinnlosen Sätzen, ebenso schränkt einen die Limitierung auf das in einem bestimmten Kontext gegebene Vokabular unnötig ein. Zudem hat im konkreten Fall Davidsons formallogische Orientierung den Annotierenden Probleme gemacht. Des Weiteren haben wir darauf verzichtet, die

Input-Sequenzen so zu manipulieren, dass das LLM explizit dazu aufgefordert wird, einer bestimmten Sprachtheorie entsprechend zu handeln. Bei den hier durchgeführten Experimenten ging es uns nur darum, wie die LLMs ohne zusätzliche Informationen oder Prompt Engineering Strategien auf die ›Gelegenheitssätze‹ ›reagieren‹.[6] Zudem sollte im Weiteren untersucht werden, inwieweit andere Kontexte die Testdatenerstellung beeinflussen und ob beispielsweise ein Korpus von nicht-fiktionalen Texten andere Vorgehensweisen nötig macht.

Fußnoten

1. https://spacy.io/
2. https://openai.com/index/hello-gpt-4o/
3. Siehe: https://github.com/pagelj/dhd2025 Es muss hier darauf hingewiesen werden, dass wir nur eine einzige Prompt-Formulierung benutzt haben und nicht getestet haben, ob andere Prompt-Formulierungen andere/bessere Resultate erbringen. Entsprechend jüngerer Studien zum Einfluss von Änderungen im Prompt Design auf die Performance der jeweiligen Modell-Prompt-Kombination (vgl. Mizrahi 2024) wäre unser Testdatensatz in weiterer Folge auf eine kontrollierbare Art und Weise zu variieren, auf deren Basis dann evaluiert werden könnte, inwiefern diese Variationen den Grad der Entsprechung des ›Sprachverhaltens‹ eines LLMs beeinflussen.
4. Ein Kappa-Wert von 0 signalisiert, dass die Übereinstimmungen/Nicht-Übereinstimmungen in den Annotationen auch zufällig entstanden sein könnten, ein negativer Wert, dass die Nicht-Übereinstimmungen höher ausfallen als per Zufall erwartet und ein positiver Wert, dass die Übereinstimmungen höher ausfallen als per Zufall erwartet, mit einem Minimal- und Maximalwert von -1 und +1. Ein Wert von +1 entsteht, wenn alle Annotator:innen in allen Fällen das gleiche Label vergeben.
5. https://www.anthropic.com/news/claude-3-5-sonnet
6. Zum Einsatz von In-Contex-Learning und Prompt Engineering in den CLS sowie die dafür relevante NLP- und LLM-Forschung siehe zum Beispiel: Pagel/Pichler/Reiter 2024 und Hicke/Bizzoni/Feldkamp/Kristensen-McLachlan 2024.

Bibliographie

Beizer, Boris. 1995. *Black-box testing: techniques for functional testing of software and systems*. New York: Wiley.

Davidson, Donald. 2009. “Radical Interpretation” In *The Essential Davidson*. Oxford: Clarendon Press, 184-195.

Davidson, Donald. 2001. “Epistemology Externalized”. In *Subjective, intersubjective, objective*. Oxford: New York: Clarendon Press; Oxford University Press, 193-204.

Fleiss, Joseph L. 1971. “Measuring Nominal Scale Agreement Among Many Raters” In *Psychological Bulletin* 76.5: 378–382.

Gengnagel, Tessa, Fotis Jannidis, Rabea Kleymann, Julian Schröter, und Heike Zinsmeister. 2024. “Bedeutung in Zeiten großer Sprachmodelle”. In *10. Tagung des Verbands Digital Humanities im deutschsprachigen Raum, DHd 2024*, herausgegeben von Joëlle Weis, Estelle Bunout, Thomas Haider. https://doi.org/10.5281/ZENODO.10698308 .

Gerstorfer, Dominik und Evelyn Gius. 2023. "Konflikte als Theorie, Modell und Text – Ein kategorientheoretischer Zugang zur Operationalisierung von Konflikten". In *9. Tagung des Verbands Digital Humanities im deutschsprachigen Raum, DHd 2023*, herausgegeben von Busch, Anna und Peer Trilcke, 189. DOI: 10.5281/ZENODO.7715293 .

Hicke, Rebecca M. M., Bizzoni, Yuri, Feldkamp, Pascale, Kristensen-McLachlan, Ross Deans. 2024. „Says Who? Effective Zero-Shot Annotation of Focalization“. arXiv. http://arxiv.org/abs/2409.11390 .

Krautter, Benjamin, Axel Pichler, und Nils Reiter. 2023. „Operationalisierung“. *Working Paper 2 der Zeitschrift für digitale Geisteswissenschaften*. Zeitschrift für digitale Geisteswissenschaften – ZfdG. https://doi.org/10.17175/WP_2023_010 .

Mizrahi, Moran, Kaplan, Guy, Malkin, Dan, Dror, Rotem, Shahaf, Dafna, Stanovsky, Gabriel. 2024. „State of What Art? A Call for Multi-Prompt LLM Evaluation“ . In: Transactions of the Association for Computational Linguistics , 12:933–949. https://aclanthology.org/2024.tacl-1.52 .

Pagel, Janis, Pichler, Axel, Reiter. 2024. „Evaluating In-Context Learning for Computational Literary Studies: A Case Study Based on the Automatic Recognition of Knowledge Transfer in German Drama“ . In Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024) , 1–10, https://aclanthology.org/2024.latechclfl-1.1 .

Scheffel, Michael. 2002. “Das Urteil – Eine Erzählung ohne ‘geraden, zusammenhängenden, verfolgabren Sinn’?”. In *Kafkas „Urteil“ und die Literaturtheorie: zehn Modellanalysen*, herausgegeben von Oliver Jahraus und Stefan Neuhaus, 59-77. Stuttgart: Reclam.

Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, u. a. 2023. „Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models“. arXiv. http://arxiv.org/abs/2206.04615 .