SIENA: Sprachmodellbasierte Identifikation und Extraktion von Nutzeranforderungen

Eine innovative Methode zur automatisierten Anforderungsanalyse

Dorian Zwanzig[1], Anja Kahl[2] und Prof. Dr.-Ing. Ute Dietrich[3]

Abstract: In diesem Artikel wird eine innovative Methode zur Identifikation und Extraktion von Nutzeranforderungen aus natürlichsprachlichen Quellen präsentiert. Als Ausgangsbasis dient uns eine Fallstudie zur Entwicklung einer Fachanwendung für Physiotherapeuten. Die Methode nutzt OpenAI's Generative Pretrained Transformer (GPT) Modelle und deren Fähigkeit zur Verarbeitung natürlicher Sprache. Eine quantitative Analyse wurde durchgeführt, um die Wirksamkeit dieser Sprachmodelle bei der Anforderungsanalyse zu bewerten. Die Ergebnisse zeigen deutlich, dass die verwendeten GPT-Modelle eine effektive und kostengünstige Unterstützung bei der Anforderungsanalyse sein können.

Keywords: Anforderungsanalyse, Nutzeranforderungen, Künstliche Intelligenz, Sprachmodell, ChatGPT, Large Language Model, Natürliche Sprache

1 Einführung

Die Identifikation und Extraktion von Nutzeranforderungen stellt einen kritischen und zeitaufwendigen Prozess im Anforderungsmanagement dar, der häufig von speziell geschulten Fachexperten durchgeführt wird und mit erheblichem betriebswirtschaftlichem Aufwand verbunden ist. Ohne eine klare, eindeutige Kenntnis der zu erreichenden Zielstellung wird das Ergebnis an den Bedürfnissen der Stakeholder vorbeigehen und mühevoll und aufwendig nachgebessert werden müssen. In den letzten Jahren haben Large Language Models (LLM) bemerkenswerte Fortschritte in der Analyse und Interpretation von textbasierten Inhalten gezeigt, wie etwa in der Sentiment-Analyse, der Erstellung von Zusammenfassungen oder der Erweiterung von Texten. Angesichts dieser Entwicklungen liegt der Fokus dieser Studie auf einer Untersuchung, in wie weit LLM, insbesondere GPT-3.5 und GPT-4, eingesetzt werden können, um den manuellen Aufwand im Bereich

[1] Hochschule für Technik und Wirtschaft (HTW) Berlin, Fachbereich 4 Informatik, Kommunikation und Wirtschaft – Studiengang Wirtschaftsingenieurwesen, Wilhelmienhofstraße 75A, 12459 Berlin, dorian.zwanzig@htw-berlin.de
[2] Hochschule für Technik und Wirtschaft (HTW) Berlin, Fachbereich 4 Informatik, Kommunikation und Wirtschaft – Studiengang Wirtschaftsingenieurwesen, Wilhelmienhofstraße 75A, 12459 Berlin, anja.kahl@student.htw-berlin.de
[3] Hochschule für Technik und Wirtschaft (HTW) Berlin, Fachbereich 4 Informatik, Kommunikation und Wirtschaft – Studiengang Wirtschaftsingenieurwesen, Wilhelmienhofstraße 75A, 12459 Berlin, ute.dietrich@htw-berlin.de

der Anforderungserhebung zu reduzieren und diesen Prozess zumindest teilweise zu automatisieren.

Die zentrale Forschungsfrage lautet daher: "Inwiefern können Large Language Models, wie GPT-3.5 und GPT-4, effektiv zur Identifikation und Extraktion von Nutzeranforderungen aus natürlichsprachlichen Quellen eingesetzt werden?"

Um diese Frage zu beantworten, wird ein Entwicklungsprojekt für eine Anwendung für Physiotherapeuten als Fallstudie herangezogen. Im Rahmen dieses Projekts wurden semi-strukturierte Interviews mit potenziellen NutzerInnen durchgeführt, um das Nutzerumfeld zu verstehen und Lösungsideen zu sammeln. Die in den Interviews geäußerten Wünsche bezüglich der bereitzustellenden Funktionalitäten und Inhalte wurden vom Projektteam analysiert und in Nutzeranforderungen übersetzt, die als Baseline für den Vergleich in dieser Studie dienen.

Nach Anonymisierung und Freigabe durch die InterviewpartnerInnen wurden relevante Abschnitte aus den Interviews extrahiert und über die OpenAI-API mit den Modellen GPT3.5-turbo-0301 und GPT-4-0314 verarbeitet. Anschließend wurden die extrahierten Anforderungen hinsichtlich ihrer Relevanz, Übereinstimmung mit der Baseline und Qualität analysiert und strukturiert. Abschließend wurden beide Modelle hinsichtlich ihrer Leistungsfähigkeit verglichen um festzustellen, ob und inwieweit LLM zur Identifikation, Extraktion und auch Strukturierung von Nutzeranforderungen beitragen können.

2 Stand der Technik

Natural Language Processing (NLP) im Bereich des Anforderungsmanagement, im Englischen des Requirement Engineerings (RE), ist ein aktives und sich entwickelndes Forschungsfeld. Bis 2019 wurden 404 Studien zu diesem Thema durchgeführt, von denen 67 % Lösungsvorschläge enthielten. Dabei bezogen sich 59 % auf die Analyse und Erhebung von Anforderungen [Zh21]. Aufgrund neuer Techniken wächst das Interesse der RE-Forscher in diesem Bereich. [FZA21] KI-basierte Tools verwenden NLP und Machine Learning (ML) hauptsächlich zur Anforderungsklassifikation, um informelle Texte in formell strukturierte Anforderungen umzuwandeln und zur Anforderungserhebung. [My21; DN20; Be22; Su19]

Aktuelle Ansätze zur Anforderungsextraktion befassen sich mit der Verwendung des BERT-Modells für RE, ChatGPT als Vertreter für Large Language Models und der Entwicklung einer Softwarelösung zur Identifikation von Wissen aus kausalen Zusammenhängen [Zh23; AM21; Fi20]. Hauptfragestellungen im Bereich RE mithilfe von NLP beinhalten die Automatisierung der Anforderungsklassifikation und die automatische Extraktion kausaler Beziehungen aus natürlicher Sprache. Zusätzlich sollen Fragen zur automatischen Generierung von Anforderungen aus Applikations-Reviews und Nutzerrezensionen inklusive der Erkennung und Vermeidung von Mehrdeutigkeiten und /oder Widersprüchlichkeiten beantwortet, Probleme bei der Erstellung formeller

Anforderungen sowie die Beeinflussung der Forschung und Praxis des NLP für das RE durch generative LLMs in den Betrachtungsfokus gerückt werden. [Fi20; DN20; Su19; AM21; Be22]

Aktuell konzentriert sich die durch NLP unterstützte Extraktion von Anforderungen hauptsächlich auf nutzergenerierte Inhalte und Anforderungsspezifikationen. Lediglich vier der bis 2019 veröffentlichten Studien beinhalteten Interview-Transkripte als Eingabedokumente. [Zh21] Die Hauptprobleme, die weitere Forschung erfordern, sind Informalität, Individualität und die in den Texten enthaltenen Rechtschreib- und Grammatikfehler; das Herausfiltern ausschließlich der für den Entwickler relevanten Informationen und der teilweise fehlende Bestand an Datensätzen, um Entwicklungen testen zu können. Insbesondere Mehrdeutigkeiten und der Gebrauch von Synonymen und Homonymen in Aussagen führen zu geringerer Leistung und Fehlern, was einen hohen Aufwand an Übersetzung in formelle Strukturen erfordert. [Be22; Su19; Z23; AM21] Der leistungsstarke ChatGPT im Bereich NLP weist nur begrenztes Wissen im Bereich RE auf. Die dazugehörige empirische Evaluation ergab jedoch, dass in der Nutzung von LLMs für Anforderungsabrufe Potenzial liegt. [Zh23] Eine grundlegende Hürde besteht aktuell darin, dass für alle Sprachmodelle noch menschliche Überwachung erforderlich ist. [Be22]

Angesichts dieser Herausforderungen und Möglichkeiten sollte zukünftige Forschung darauf abzielen, die Leistungsfähigkeit und Anwendbarkeit von NLP- und LLM-Techniken im Bereich RE weiter zu verbessern. Dazu gehört die Untersuchung neuer Ansätze zur Reduzierung von Mehrdeutigkeiten, die Optimierung von Eingabedaten und die Anpassung von Sprachmodellen an RE-spezifische Aufgaben. Darüber hinaus sollte die Forschung auch die Integration verschiedener NLP- und LLM-Tools in bestehende RE-Methoden und -Prozesse in Betracht ziehen, um Synergien zu nutzen und die Effektivität der Anforderungserhebung und -analyse insgesamt zu erhöhen.

3 Methode

Die semi-strukturierten Interviews wurden anhand eines thematisch abgestimmten Leitfadens durchgeführt, der drei Hauptbereiche fokussierte: das Nutzerumfeld inklusive typischer Tätigkeiten und beruflicher Herausforderungen, den Wissenserwerb im beruflichen Kontext mit besonderem Fokus auf medizinischen Leitlinien sowie Ideation d.h. die Entwicklung von Lösungsansätzen für eine digitale Wissensbereitstellung. Für die hier diskutierte Studie ist vor allem der dritte Teil relevant. Ausgewählte Interviewausschnitte zeichneten sich durch ihre Anforderungsdichte aus, das heißt, sie enthielten eine Vielzahl von Anforderungen und konnten auch unabhängig von weiterem Kontext verarbeitet werden. Drei Interviewpartner gaben ihre Zustimmung zur Nutzung ihrer Interviews in dieser Studie und durch Drittsysteme, hier OpenAI, wodurch eine Datenbasis von 16 Interviewabschnitten mit insgesamt etwa 5.000 Wörtern entstand. Jeder Abschnitt umfasste typischerweise ein Frage-Antwort-Paar. Insgesamt wurden 46

Nutzeranforderungen manuell aus diesen Ausschnitten extrahiert, die als Baseline für die Beurteilung der Leistungsfähigkeit der eingesetzten Sprachmodelle dienen.

Die Interviewausschnitte enthielten primär Softwareanforderungen aus der NutzerInnen-Perspektive. Es sollte eine mobile Anwendung entwickelt werden, die es PhysiotherapeuthInnen ermöglicht, innerhalb kurzer Zeit auf evidenzbasierte Behandlungsempfehlungen zurückzugreifen. Dabei wurden in der manuellen Analyse unter anderem folgende Anforderungen identifiziert:

- Die Anwendung muss die Eingabe von Diagnosen zur Suche nach Behandlungen ermöglichen.
- Die Anwendung soll eine kurze Übersicht über die Diagnose und Behandlung bereitstellen.
- Die Anwendung soll Informationen als Fließtext bereitstellen.

In einem iterativen Prozess wurde ein ausführlicher Prompt entwickelt, der als Eingabe für die Modelle diente. Bei einem Prompt handelt es sich im Grunde um eine Anweisung in Textform an den ChatGPT oder ein vergleichbares KI-System. Der im Projekt entwickelte Prompt bestand aus zwei Teilen: einem zur Definition des Systems und seiner Aufgabe und einem zur Bereitstellung des Interviewausschnitts. Der erste Teil definierte den Kontext der Aufgabe, hier die Anwendungsentwicklung für Physiotherapeuten, die Rolle des Systems als Expertensystem, dessen Aufgabe, die auf eine Identifikation und Extraktion von strukturierten Anforderungen fokussiert sowie das Ausgabeformat. Der zweite Teil enthielt den jeweiligen Interviewausschnitt.

Im folgenden ist der Systemprompt dargelegt: „Du bist ein System, das Anforderungen in natürlich sprachlichen Interviews identifiziert und extrahiert. Dir wird vom """user""" ein Interviewausschnitt präsentiert. In dem Interview geht es um die Entwicklung einer Anwendung zur Bereitstellung von medizinischem Fachwissen für PhysiotherapeutInnen. Der Ausschnitt startet mit einer Frage durch einen Interviewer "I". Der Interviewer stellt eine oder mehrere Fragen an den Interviewpartner "B". Der Interviewpartner "B" antwortet auf die Frage.Du analysierst den Interviewausschnitt und versuchst, direkte Hinweise auf die vom Interviewpartner "B" gewünschten Funktionen der Anwendung zu identifizieren. Ignoriere Anforderungen, die nur aus der Aussage des Interviewers "I" resultieren. Du gibst anschließend eine Liste von Anforderungen aus, die du aus dem Interviewausschnitt extrahiert hast. Die Anforderungen sollen in funktionale und nicht-funktionale Anforderungen unterteilt werden. Du kannst auch eine Anforderung als unklar markieren, wenn du dir nicht sicher bist, ob sie eine Anforderung ist. Es sollen keine impliziten Anforderungen ausgegeben werden. Jede Anforderung soll mit einem Zitat aus dem Interviewausschnitt belegt werden. Die Anforderungen soll nach folgendem Schema formuliert werden: Das System muss/kann/soll + <Anforderung>. Gib das Ergenis deiner Analyse als CSV aus, nutze als Trennzeichen ";" und folgende Struktur: Funktional/nicht funktional/unklar; Anforderung; Beschreibung der Anforderung; Zitat aus Interviewausschnitt.“

Die Interviewausschnitte wurden sequenziell durch die Modelle verarbeitet und die generierten Ausgaben in einer Gesamtliste zusammengefasst. Die Modelle wurden lediglich hinsichtlich des Parameters "Temperature" konfiguriert, welcher den Grad der Zufälligkeit in der Generierung der Ausgaben regelt. Dieser wurde auf 0 gesetzt, um eine Reproduzierbarkeit zu gewährleisten.

Der vollständige Prompt, die Interviewausschnitte und die generierten Anforderungen sind im frei zugänglichen und wiederverwendbaren Projekt-Repository unter https://github.com/dozwa/siena-001 einsehbar und können frei für eigene Experimente verwendet werden.

Die extrahierten Anforderungen wurden in einem nachfolgenden Schritt manuell auf Relevanz überprüft und Duplikate entfernt. Anschließend wurden die generierten Anforderungen mit der Baseline abgeglichen und die Übereinstimmung quantitativ analysiert.

GPT-3.5 und GPT-4 wurden für diese Studie aufgrund ihrer einfachen Verfügbarkeit und Anwendbarkeit ausgewählt. Die leichte Zugänglichkeit macht diese Modelle besonders relevant für Anwendungen im betrieblichen Kontext. Die umfangreichen Dokumentationen ermöglichen es zudem, diese Modelle mit relativ wenig technischem Wissen zu verwenden. Hier sind lediglich grundlegende Erfahrungen in der Programmierung mit Python erforderlich. Daher sind GPT-3.5 und GPT-4 für den praktischen Einsatz besonders geeignet.

In dieser Studie wurden die statischen Varianten der Modelle, gpt-3.5-turbo-0301 und gpt-4-0314, verwendet. Diese Modelle wurden seit ihrer Veröffentlichung nicht mehr aktualisiert, um die Wiederholbarkeit der Versuche zu gewährleisten und zu verhindern, dass das Modell aus der iterativen Durchführung der Prompt-Entwicklung lernt.

Die abschließende quantitative Analyse basierte auf der Ermittlung verschiedener Werte. Diese umfassen:

- True Positives (TP): Anzahl der korrekt identifizierten relevanten Anforderungen (in der Baseline enthalten und vom System erkannt).
- False Positives (FP): Anzahl der fälschlicherweise als relevant identifizierten Anforderungen (nicht in der Baseline enthalten, aber vom System erkannt).
- False Negatives (FN): Anzahl der relevanten Anforderungen, die vom System nicht erkannt wurden (in der Baseline enthalten, aber vom System nicht erkannt).
- True Negatives (TN): Anzahl der korrekt identifizierten, irrelevanten Anforderungen (nicht in der Baseline enthalten und vom System nicht erkannt).

Anhand dieser Werte wurden folgende Kennzahlen berechnet:

- Genauigkeit (Accuracy): Anteil der korrekt klassifizierten Anforderungen (sowohl relevant als auch irrelevant) an der Gesamtzahl der Anforderungen. Berechnet durch: (TP + TN) / (TP + FP + FN + TN)
- Präzision (Precision): Anteil der korrekt identifizierten, relevanten Anforderungen an der Gesamtzahl der vom System als relevant identifizierten Anforderungen. Berechnet durch: TP / (TP + FP)
- Sensitivität (Recall): Anteil der korrekt identifizierten, relevanten Anforderungen an der Gesamtzahl der tatsächlich relevanten Anforderungen (Baseline). Berechnet durch: TP / (TP + FN)
- F1-Wert (F1-Score): Harmonisches Mittel aus Präzision und Sensitivität, das ein ausgewogenes Maß für die Leistung des Systems bietet, insbesondere bei ungleicher Verteilung der relevanten und irrelevanten Anforderungen. Berechnet durch: 2 * (Präzision * Sensitivität) / (Präzision + Sensitivität)

4 Ergebnisse

Das GPT-3.5-Modell identifizierte und extrahierte insgesamt 47 relevante Anforderungen aus den Interviews. Davon stimmten 33 (True Positive) genau oder zumindest annähernd mit den Anforderungen aus der Baseline überein. Es ist zu beachten, dass mehrere generierte Anforderungen einer Baseline-Anforderung zugeordnet sein können. Insgesamt konnte das Modell 56% der Baseline-Anforderungen identifizieren. Darüber hinaus wurden 12 relevante Anforderungen erkannt, die nicht in der Baseline enthalten waren (False Positive). Diese waren den Analysten entweder nicht aufgefallen oder wurden nicht als relevant eingestuft. Das Modell schuf insofern einen Mehrwert, als es bisher unbekannte Anforderungen aufdecken konnte. Für 15 Baseline-Anforderungen wurden keine Entsprechungen unter den generierten Anforderungen gefunden (False Negative), und eine generierte Anforderung war völlig irrelevant (True Negative). Es lässt sich feststellen, dass etwa ¾ der vom System generierten Anforderungen relevant waren, das heißt, sie wurden korrekt aus den Interviews extrahiert.

Das GPT-4-Modell identifizierte mit 58 Anforderungen deutlich mehr Anforderungen, von denen 39 direkt den Baseline-Anforderungen zugeordnet werden konnten. Die Anzahl der False Positives und False Negatives lag in etwa im Bereich des Vergleichsmodells. Ebenso vergleichbar sind Genauigkeit, Präzision und Sensitivität. Demgegenüber stehen drei irrelevante Anforderungen, die keiner Baseline-Anforderung entsprechen. Dies entspricht der dreifachen Menge an als True Negative klassifizierten Anforderungen im Vergleich zum GPT-3.5-Modell.

Kennzahl	GPT3.5	GPT4
Baseline Anforderungen	46	46
Generierte Anforderungen	47	58
True Positives	33	39
False Positives	12	13
False Negatives	15	17
True Negatives	1	3
Accuracy	56 %	58 %
Precision	73%	75%
Recall	69%	70%
F1-Score	71%	72%

Tab. 1: Vergleich GPT3.5 und GPT4

Die Interpretation der Ergebnisse wird durch zwei Faktoren erschwert: (1) Die Baseline umfasst nicht alle möglichen Anforderungen, und die Modelle haben Anforderungen identifiziert, die nicht in der Baseline enthalten sind. (2) Die Zuordnung der Anforderungen weist eine Kardinalität von 1-n auf, sodass mehrere generierte Anforderungen einer Baseline-Anforderung zugeordnet werden können. Um diese Probleme zu lösen, wurden zwei neue modellspezifische Baselines erstellt, welche um die zusätzlichen, vom System erkannten relevanten Anforderungen erweitert wurden. Zudem wurden generierte Anforderungen mit einer Kardinalität ungleich 1-1 von der Analyse ausgeschlossen.

Nach dieser Bereinigung weisen beide Modelle erwartungsgemäß eine Präzision von 100% auf, das bedeutet, dass alle generierten Anforderungen Teil der Baseline sind. Die Genauigkeit des GPT-4-Modells liegt mit 74% etwas höher als die des GPT-3.5-Modells.

Kennzahl	GPT3.5	GPT4
Baseline Anforderungen	58	59
Generierte Anforderungen	47	58
True Positives	41	43
False Positives	0	0
False Negatives	17	16
True Negatives	1	3
Accuracy	71%	74%
Precision	100%	100%
Recall	71%	73%
F1-Score	83%	84%

Tab. 2: Vergleich GPT3.5 und GPT4 mit erweiterter Baseline:

5 Diskussion

In der quantitativen Analyse wurde gezeigt, dass die beiden Modelle nur geringfügige Unterschiede bei der Erkennung von Anforderungen aufweisen. Dies erscheint zunächst kontraintuitiv, da das GPT-4-Modell deutlich komplexer ist als das GPT-3.5-Modell. Stand Mai 2022 ist die Anwendung des GPT-4-Modells etwa 30-mal teurer als die des GPT-3.5-Modells, was aus wirtschaftlicher Sicht eine Nutzung des GPT-3.5-Modells empfehlenswert erscheinen lässt. Die Kosten für die einfache Verarbeitung der 16 Interviewausschnitte beliefen sich im Juni 2023 unter Verwendung des GPT3.5 Modells auf ca. 4 Euro Cent und mit dem GPT4 Modell auf 86 Euro Cent. Der Umfang der Anfrage belief sich je Interviewausschnitt auf ca. 1000 Token (40% Systemprompt, 60% Interviewausschnitt), die Antworten umfassten durchschnittlich ca. 400 Token. Die gesamte Verarbeitung der Interviews umfasste insgesamt ca. 22.400 Token. Die tatsächlichen API-Kosten für die gesamte Studie betrugen jedoch weniger als 5 Euro. Eine wirtschaftliche Betrachtung wäre dennoch bei umfangreicheren Anforderungsanalysen sinnvoll. Würde man das hier diskutierte Vorgehen auf ganze Interviews übertragen, würde die Analyse eines ca. einstündigen Interviews mit einem Transkript von ca. 9.000 Wörten (ca. 17.000 Token) etwa 1,12 € (GPT4) bzw. 0,05 € (GPT3.5) kosten.

Prompt	Token gesamt	Kommentar	Kosten GPT3.5 je 1000 Token	Kosten GPT4 je 1000 Token
Systemprompt	6.800	40% des Interviews	0,0015 €	0,0300 €
Interviewausschnitt	17.000	1 Std., 9.000 Worte	0,0015 €	0,0300 €
Antwort	6800	40% des Interviews	0,0020 €	0,0600 €
Gesamtkosten			**0,0493 €**	**1,1220 €**

Tab. 3: Kostenkalkulation für die Analyse eines einstündigen Interviews

Generell scheint der Einsatz von Large Language Modellen, hier am Beispiel der Modelle von OpenAI, für die Identifikation und Extraktion von Nutzeranforderungen sinnvoll zu sein. Besonders hervorzuheben ist, dass die Modelle in der Lage waren, Anforderungen aufzudecken, die in der manuellen Analyse nicht korrekt herausgearbeitet wurden.

Zukünftige Untersuchungen könnten bereits von einer erweiterten Baseline ausgehen, die auch auf den Ergebnissen aller berücksichtigter Modelle und den manuell identifizierten Anforderungen basieren. Dies würde eine noch bessere Vergleichbarkeit gewährleisten.

Der in dieser Studie verwendete Prompt erwies sich als zielführend, jedoch bedeutet dies nicht, dass er optimal gestaltet ist. Es handelt sich um einen Zero-Shot-Prompt, d.h., dieser ist auf andere Aufgaben übertragbar, ohne das zusätzliche markierte Trainingsbeispiele benötigt werden. Lediglich die Aufgabe muss entsprechend beschrieben werden. Alternativ hätte man einen Multi-Shot-Prompt verwenden können, der neben der Anweisung auch konkrete Beispiele für die Lösung der Aufgabe enthält.

Während in dieser Studie ausschließlich quantitative Aspekte berücksichtigt wurden, sollten zukünftig auch qualitative Aspekte systematisch untersucht werden. Die Qualität der generierten Anforderungen variiert erheblich. Einige Anforderungen wurden in hervorragender Qualität formuliert und berücksichtigen alle relevanten Aspekte einer Anforderungsdefinition. Andere Anforderungen bestanden jedoch aus lediglich zwei Wörtern, ließen viel Interpretationsspielraum und/oder waren strukturell inkorrekt aufgebaut.

Es sollte auch beachtet werden, dass die Modelle die Anweisung hatten, ausschließlich explizit benannte Nutzerwünsche zu berücksichtigen. Die Nutzung der Modelle zur Identifikation von impliziten Anforderungen oder Basis-Anforderungen, z.B. nach dem Kano-Modell, könnte zusätzliche interessante Erkenntnisse und Einblicke liefern.

Ein weiterer diskussionswürdiger Aspekt betrifft die Auswahl der Modelle. Die getroffene Auswahl basiert auf Zugänglichkeit und Verfügbarkeit, ohne umfassende Marktrecherche. Zukünftige Untersuchungen könnten mehr Modelle, insbesondere Modelle unterschiedlicher Anbieter sowie Open-Source-Lösungen, berücksichtigen. Es wäre auch

interessant zu untersuchen, ob ein Open-Source-Modell speziell für diese Aufgabe nachtrainiert werden könnte.

6 Zusammenfassung

Die Untersuchung hat gezeigt, dass der Einsatz der betrachteten Modelle zur Identifikation und Extraktion von Nutzeranforderungen aus natürlichsprachlichen Quellen vielversprechend ist. Beide analysierten Modelle konnten einen Großteil der relevanten Anforderungen aus den Texten extrahieren, wobei das GPT-4-Modell dem weniger komplexen GPT-3.5-Modell leicht überlegen ist. Hervorgehoben werden muss, dass beide Modelle neue, relevante Anforderungen identifiziert haben. Dennoch sollte die Quantität nicht das einzige Kriterium sein, anhand dessen die Leistung gemessen wird. Eine gesonderte Betrachtung der Qualität der Anforderungen wird dringend empfohlen.

Die Autoren sind der Meinung, dass dieser Ansatz positive praktische Auswirkungen haben kann, indem er eine sinnvolle Ergänzung zur manuellen Anforderungsanalyse darstellt. Bei einer exponentiell ansteigenden Anforderungsanzahl komplexer Produkte ist eine manuelle Erhebung schlicht kaum noch zu gewährleisten. Möglicherweise könnten die Modelle auch bei der Identifikation von impliziten Anforderungen nützlich sein. Wirtschaftliche Aspekte sollten bei der Auswahl der Modelle berücksichtigt werden, da die hier betrachteten Modelle signifikant unterschiedliche Kosten verursachen.

Die Leistungsfähigkeit der Modelle könnte durch einen umfangreicheren Modellvergleich weiter objektiviert werden. Durch gezieltes Nach-Training der Modelle oder eine Optimierung der Prompts könnte ihre Leistungsfähigkeit zusätzlich verbessert werden.

Literaturverzeichnis

[AM21] Araújo, A.; Marcacini, R.: RE-BERT: Automatic Extraction of Software Requirements from App Reviews using BERT Language Model. In (Association for Computing Machinery): SAC'21: Proceedings of the 36th Annual ACM Symposium on Applied Computing, S. 1321-1327, 2021.

[Be22] Bertram, V. et.al.: Neural Language Models and Few Shot Learning for Systematic Requirements Processing in MDSE. In (Association for Computing Machinery): SLE 2022: Proceedings oft he 15th ACm SIGPLAN International Conference on Software Language Engineering. Auckland, S. 260-265, 2022.

[DN20] Dalpiaz, F.; Niu, N.: Requirements Engineering in the Days of Artificial Intelligence. IEEE Software, Band 37, Ausgabe 4, S. 7-11, 2020.

[Fi20] Fischbach, J. et.al.: Towards Causality Extraction from Requirements. In (IEEE): 2020 IEEE 28th International Requirments Engineering Conference (RE). Zürich, S. 388-393, 2020.

[FZA21] Ferrari, A.; Zhao L., Alhoshan W.: NLP for Requirements Engineering: Tasks, Techniques, Tools, and Technologies. In (IEEE): 2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion). Madrid, S. 322-323., 2021.

[My21] Myllynen, S. et.al.: Developing and Implementing Artificial Intelligence-Based Classifier for Requirments Engineering. ASME J of Nuclear Rad Sci. Band 7, Ausgabe 3, S. 1, 2021.

[Su19] Surana, C. et.al.: Intelligent Chatbot for Requirements Elicitation and Classification. In (IEEE): 2019 4th International Conference on Recent Traneds on Electronics, Information, Communication & Technology (RTEICT). Bangalore, S. 866-870, 2019.

[Zh21] Zhao, L. et al.: Natural Language Processing for Requirements Engineering: A Systematic Mapping Study. ACM Computing Survey, Band 54, Ausgabe 3, Art. 55, S. 1-41, 2021.

[Zh23] Zhang, J. et.al.: A Preliminary Evaluation of ChatGPT in Requirements Information Retrieval. 2023.