

Aus der Klinik für Mund-, Kiefer- und Gesichtschirurgie, Plastische Operationen
der Universitätsmedizin der Johannes Gutenberg-Universität Mainz

Künstliche Intelligenz in der medizinischen Dokumentation und
Patientenkommunikation – Eine vergleichende Studie zur Lesbarkeit,
Verständlichkeit und Patientenakzeptanz von ChatGPT-generierten
Röntgenberichten

Inauguraldissertation
zur Erlangung des Doktorgrades der
Zahnmedizin
der Universitätsmedizin
der Johannes Gutenberg-Universität Mainz

Vorgelegt von

Annika Sophie Bertsch
aus Groß-Umstadt

Mainz 2025

Wissenschaftlicher Vorstand:	Univ.-Prof. Dr. Philipp Drees
1. Gutachter:	PD Dr. Dr. Daniel Thiem
2. Gutachter:	PD Dr. Alexander Tobias Ziebart
Tag der Promotion:	01. Oktober 2025
Nachnutzungslizenz:	CC-BY-4.0

Inhaltsverzeichnis

1	<u>EINLEITUNG</u>	2
1.1	DAS PATIENTENRECHTEGESETZ UND SEINE KONSEQUENZEN	2
1.2	KÜNSTLICHE INTELLIGENZEN IN DER KLINISCHEN ANWENDUNG	2
1.3	CHATGPT	4
1.3.1	ENTWICKLUNG DER LARGE LANGUAGE MODELS ZUM TRANSFORMER-MODEL	4
1.3.2	ALGORITHMUS-INDUZIERTER DISKRIMINIERUNG	6
1.3.3	HALLUZINATIONEN DER LARGE LANGUAGE MODELS	6
1.3.4	PROMPT-ENGINEERING	7
1.4	MEDIZINISCHE KOMPETENZ UND AKTUALITÄT DER GPT-VERSIONEN	7
1.4.1	MÖGLICHKEITEN ZUR EFFIZIENZSTEIGERUNG FÜR DEN KLINISCHEN ALLTAG	8
1.5	LESBARKEIT PATIENTENORIENTIERTER INFORMATIONSSCHRIFTEN	9
1.5.1	ÜBERWINDUNG VON SPRACHBARRIEREN DURCH DIE ÜBERSETZUNGSKOMPETENZ DER KI	10
1.5.2	PATIENTENVERTRAUEN IN KÜNSTLICHE INTELLIGENZ	11
1.6	ZIELSETZUNG	11
2	<u>MATERIAL UND METHODEN</u>	14
2.1	DATENERHEBUNG	14
2.2	KI-TEXTGENERIERUNG	16
2.2.1	MODELLPARAMETER VON CHATGPT	16
2.2.2	DATENTRANSFORMATION	17
2.2.3	PROMPT UND KI-TEXTGENERIERUNG	17
2.3	DATENAUSWERTUNG	17
2.3.1	ANALYSE DER STUDIERENDEN-FORMULIERTEN RÖNTGENBERICHTE	18
2.3.2	LESBARKEITSINDIZES	18
2.3.2.1	Flesch Reading Ease-Score	18
2.3.2.2	Lesbarkeitsindex	19
2.3.3	ANALYSE DER TEXTGENAUIGKEIT MITTELS BERT-SCORE	20
2.4	VEREINFACHUNG DER KI-GENERIERTEN RÖNTGENBERICHTE	21
2.5	TEXTEVALUATION DURCH PATIENTEN	22
2.5.1	PATIENTENKOLLEKTIV	22
2.5.2	FRAGEBÖGEN	22
2.6	POWER ANALYSE UND FALLZAHLBERECHNUNG	24
2.7	STATISTISCHE ANALYSE	24
2.8	ETHIKVOTUM	25
3	<u>ERGEBNISSE</u>	26
3.1	RÖNTGENBERICHTE - CHATGPT VS. STUDIERENDE	26
3.1.1	ZEITAUFWAND ZUR ERSTELLUNG DER RÖNTGENBERICHTE	26
3.1.2	LESBARKEIT DER RÖNTGENBERICHTE	26
3.1.3	ÄHNLICHKEIT UND INFORMATIONSGENAUIGKEIT DER RÖNTGENBERICHTE	28
3.1.4	QUANTITATIVE ANALYSE DER SPRACHLICHEN MERKMALE UND DER FEHLERQUOTE	29
3.2	AUSWERTUNG DER VON CHATGPT VEREINFACHTEN RÖNTGENBERICHTE	31
3.2.1	LESBARKEIT DER VEREINFACHTEN RÖNTGENBERICHTE	31
3.2.2	QUANTITATIVE ANALYSE DER SPRACHLICHEN MERKMALE DER VEREINFACHTEN RÖNTGENBERICHTE	33
3.3	PATIENTENEVALUATION DER KI-GENERIERTEN RÖNTGENBERICHTE	34

3.3.1	KLARHEIT UND DEUTLICHKEIT DER RÖNTGENBERICHTE	34
3.3.2	STRUKTUR UND INFORMATIONSGEHALT DER RÖNTGENBERICHTE	35
3.3.3	TON UND EMPATHIE DES RÖNTGENBERICHTS	36
3.3.4	ANLEITUNG UND MOTIVATION DER RÖNTGENBERICHTE	36
3.3.5	ZUKÜNFTIGE PERSPEKTIVE DER RÖNTGENBERICHTE	38
4	<u>DISKUSSION</u>	39
4.1	EFFIZIENZSTEIGERUNG DURCH KI-GESTÜTZTE DOKUMENTATION	40
4.2	VERGLEICH DER TEXTQUALITÄT UND LESBARKEIT	41
4.3	INFORMATIONSGENAUIGKEIT UND VOLLSTÄNDIGKEIT	42
4.3.1	EINFLUSS DES FORMATIERUNGSDESIGNS AUF DIE INFORMATIONSDICHTEN VON KI-GENERIERTEN BERICHTEN	43
4.3.2	ANALYSE UND KORREKTUR VON FEHLINTERPRETATIONEN IN KI-GENERIERTEN BERICHTEN	44
4.4	PATIENTENPERSPEKTIVE UND -VERTRAUEN	45
4.4.1	VERSTÄNDLICHKEIT UND INFORMATIONSGEHALT DER KI-GENERIERTEN BERICHTE	45
4.4.2	EMPATHIE UND TON	47
4.5	SICHERHEITSSTRUKTUREN IN DER ZUSAMMENARBEIT MIT KI	49
4.6	PERSPEKTIVEN DER IMPLEMENTIERUNG	50
5	<u>ZUSAMMENFASSUNG</u>	53
6	<u>LITERATURVERZEICHNIS</u>	55
7	<u>ABBILDUNGSVERZEICHNIS</u>	62
8	<u>TABELLENVERZEICHNIS</u>	65
9	<u>ANHANG</u>	66
9.1	RÖNTGENBERICHTE DER PATIENTENEVALUATION	66
9.1.1	PROMPT 1:	66
9.1.2	PROMPT 2:	66
9.1.3	PROMPT 3:	67
9.2	EUROPÄISCHES ZAHNSCHEMA	68
9.3	BEISPIELE DER EXCEL-TABELLE	69
9.4	DANKSAGUNG	71
10	<u>TABELLARISCHER LEBENS LAUF</u>	72

Abkürzungsverzeichnis

AWMF	Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e.V.
BERT	Bidirectional Encoder Representations from Transformer
BGB	Bürgerliches Gesetzbuch
CEC	Constant Error Carousel
DL	Deep Learning
DSGVO	Datenschutz-Grundverordnung
EKG	Elektrokardiogramm
FDI	Fédération Dentaire Internationale
FRE	Flesch Reading Ease-Score
GPT	Generative Pretrained Transformer
GRC	Charta der Grundrechte der Europäischen Union, Grundrechtechart
JMLE	Japanese Medical Licency Examination
KI	Künstliche Intelligenz
LIX	Lesbarkeitsindex
LLM	Large Language Model
LSTM	Long Short-Term Memory
OPG	Orthopantomogramm
RNN	Recurrent Neural Network
USMLE	United States Medical Licency Examination

1 Einleitung

1.1 Das Patientenrechtegesetz und seine Konsequenzen

Mit dem Inkrafttreten des „Patientenrechtegesetzes“ § 630f-g BGB im Februar 2013 wurde die Dokumentationspflicht eines Behandlers im medizinischen Kontext erstmals im Rahmen des bürgerlichen Gesetzes geregelt [1]. Während die Verabschiedung dieses Gesetzes eine erhöhte Transparenz für den Patienten darstellt, resultiert für den behandelnden Arzt eine gesetzliche Pflicht zur präzisen Niederschrift der Patientenakte. Als Konsequenz ergibt sich ein erhöhter Zeitaufwand zur vollständigen und ausführlichen Dokumentation, der durchschnittlich und mit steigender Tendenz vier Stunden oder 44% der täglichen Arbeitszeit eines Arztes einnimmt. Damit einher gehen erhöhte Dokumentationskosten, die sich auf 21% des gesamten Personalaufwands belaufen [2]. Auch wenn eine ausführliche Patientenakte das Vertrauensverhältnis zwischen einem Patienten und seinem Behandler unterstützen soll, zeigt der Alltag an deutschen Kliniken einen verminderten Patientenkontakt aufgrund des Zeitmangels, der unter anderem das Resultat der Dokumentationspflicht ist. Als Folge weicht das Berufsbild des Arztes immer weiter davon ab, was sich Berufseinsteigende darunter vorgestellt haben. Während die Antwort der meisten Medizinstudierende auf die Frage nach der Motivation, ein Medizinstudium und somit die Karriere eines Arztes einzuschlagen, das Ziel „mit Menschen zu arbeiten“ ist [3], zeichnet die Realität ein anderes Bild. So sagen 27% der Ärzte, dass die Dokumentationslast zu Demotivation und Unzufriedenheit in ihrem Beruf führt [2]. In Zeiten des Ärztemangels, der aufgrund des demographischen Wandels nicht mehr zu vernachlässigen ist, gilt es, den medizinischen Beruf attraktiver zu gestalten und die Gründe für einen frühzeitigen Berufsausstieg zu hinterfragen [4]. Es stellt sich hier die Frage nach Lösungsansätzen, die den ärztlichen Berufsalltag erleichtern und dem Problem des Zeitmangels und des bürokratischen Aufwands Abhilfe verschaffen.

1.2 Künstliche Intelligenzen in der klinischen Anwendung

Die Nutzung von künstlichen Intelligenzen (KI), insbesondere dem Deep Learning (DL), stellt in der Medizin längst kein Zukunftsszenario mehr dar. Vielmehr wird heutzutage an vielen Stellen im Gesundheitswesen KI zur Effizienzsteigerung eingesetzt. Untersuchungen zeigen, dass KI unter anderem in der Hautkrebsvorsorge [5], in der Analyse von Elektrokardiogrammen (EKGs) [6] oder der Untersuchung von pathologischen Präparaten [7] genau so gute, teilweise sogar bessere Diagnosen und Vorhersagen trifft, wie das hochqualifizierte ärztliche Gesundheitspersonal. Besonders auf dem Gebiet der radiologischen Bilderkennung werden bereits KI-Systeme erfolgreich eingesetzt und können so Radiologen entlasten [8].

Dennoch müssen vor der Implementierung von Künstlicher Intelligenz ethische und rechtliche Aspekte Beachtung finden, weshalb die *Hochrangige Expertengruppe für Künstliche Intelligenz* (HLEG), eingesetzt von der Europäischen Kommission, 2019 eine Ethikleitlinie für eine vertrauenswürdige künstliche Intelligenz veröffentlicht hat [9]. Aus dieser Leitlinie lässt sich die Forderung nach der Einhaltung geltender Rechts- und Verwaltungsvorschriften, die Achtung ethischer Grundsätze, sowie die Robustheit aus technischer Sicht als auch unter Berücksichtigung des sozialen Umfelds ableiten. Darüber hinaus ist das Vertrauen in den Behandler in der Medizin ein vulnerables Thema und daher muss der zum Ergebnis führende Prozess transparent gestaltet werden. Die so genannte „Black-Box-Problematik“, welche die für uns Menschen unvollständige Plausibilität der Ergebnisfindung beschreibt [9], ist nur durch eine Aufschlüsselung und somit Interpretierbarkeit der Algorithmen hinter der KI lösbar [10, 11].

Die unzureichende Nachvollziehbarkeit in der Erstellung des Output birgt unkalkulierbare Risiken, die aktuell noch im Zentrum des juristischen Diskurses um den Einsatz von KI in der Medizin und Pflege stehen [12] und die Überwachungspflicht und Kontrollmaßnahmen gegenüber der KI erhöhen sollten [13]. Die Herausforderung, die die klinische Nutzung von KI mit sich bringt, ist somit die Fragestellung nach der juristischen und strafrechtlichen Verantwortung und so wird laut Beck et al. der „letztentscheidende Mensch (...) zum symbolischen Haftungsknecht“ [12], welcher in diesem Fall der medizinische Behandler darstellt. Dieser ist laut § 630a BGB-Absatz 2 dazu verpflichtet, seine „Behandlung nach den zum Zeitpunkt der Behandlung bestehenden, allgemein anerkannten fachlichen Standards“ zu richten. Einen Maßstab hierfür bieten den behandelnden Ärzten die aktuellen AWMF-Leitlinien (Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e.V.), welche jedoch derzeit kaum die Empfehlung zur Anwendung von KI beinhalten [13]. Schließlich ist die tatsächliche Frage nach der Verantwortung und somit der Risikotragung noch nicht vollends geklärt. Vielmehr wird zwischen der durch den Patienten, die durch den Behandelnden und jene durch den Hersteller der KI unterschieden [13].

Obwohl der Einsatz von KI in der Medizin folglich noch in den sprichwörtlichen Kinderschuhen steckt und die Implementierung autonomer Algorithmen im klinischen Bereich vor verschiedenen Herausforderungen steht, wird das Potenzial von sprach- und dialoggesteuerter KI, also Large Language Models (LLM), bei repetitiven Aufgaben wie der Dokumentation und Textproduktion bereits deutlich erkennbar [14, 15]. Eine solche KI ist darauf ausgerichtet, die menschliche Sprache nachzuahmen und auf Basis der zur Verfügung gestellten und erlernten Informationen auf eine menschliche Art zu kommunizieren [16].

1.3 ChatGPT

ChatGPT (Generative Pre-trained Transformer), ein von OpenAI erstmals im Jahr 2018 in der Vorgängerversion GPT-1 veröffentlichtes Large Language Model, ist zu genau solchen Leistungen imstande. Seitdem wurde es kontinuierlich weiterentwickelt, bis im November 2022 die öffentlich zugängliche Version ChatGPT-3.5 eingeführt wurde [17]. ChatGPT-3.5 ist eine neuere Version, die speziell für die Interaktion in einem Chatformat optimiert wurde und mit einer benutzerfreundlichen Oberfläche auf Benutzeranfragen reagieren sowie mit diesen interagieren kann [18]. Darüber hinaus kann der Bot kreative Texte wie Geschichten, Gedichte oder sogar ganze Skripte erstellen, die ein hohes Maß von Kohärenz aufweisen und menschliche Sprache wirkungsvoll imitieren [19, 20]. Diese von der KI erstellten Texte sind zu einem sehr großen Teil frei von Rechtschreib- und Grammatikfehlern.

Die Technologie hinter diesen Modellen basiert auf dem sogenannten Deep Learning, welches auf einer Struktur aus neuronalen Netzwerken aufbaut. Die hierüber dargestellten Algorithmen ahmen die Funktionsweise des menschlichen Gehirns nach und sind somit zur komplexen Mustererkennung fähig [21]. Dieses „tiefe“ Lernen bezieht sich auf die Anzahl der Schichten, in denen die Daten erlernt und somit auch kombiniert werden, worüber eine komplexere und abstraktere Reproduktion der erlernten Informationen ermöglicht wird [22].

1.3.1 Entwicklung der Large Language Models zum Transformer-Model

ChatGPT baut speziell auf einem Ansatz namens Transformer auf, der es dem Model ermöglicht, große Mengen von Textdaten zu verstehen, wiederkehrende Muster in diesen Daten zu erkennen und auf dieser Basis neue Texte zu generieren, die den gelernten Texten entsprechen [23, 24]. Dieser Transformer wurde 2017 in der Publikation „Attention is all you need“ von Vaswani et al. vorgestellt und ist ein neuartiges Model für die Sequenztransduktion, das vollständig auf Aufmerksamkeitsmechanismen (self-Attention) basiert und die in Encoder-Decoder Architekturen üblicherweise verwendeten rekurrenten Schichten ersetzt [25, 26].

Die Encoder-Decoder-Architektur ist ein System, das oft in der Verarbeitung von Sequenz-zu-Sequenz-Aufgaben verwendet wird, wie zum Beispiel beim maschinellen Übersetzen von einer Sprache in eine andere [27]. Der Encoder verarbeitet die Eingabesequenz so, dass die Informationen als Basis für die Generierung der Ausgabesequenz dienen können. Nachdem der Encoder die Eingabesequenz verarbeitet hat, nimmt der Decoder diese und generiert schrittweise die Ausgabesequenz [28].

Traditionell bestehen die Architekturen von Encodern und Decodern häufig aus mehreren rekurrenten Schichten (Recurrent Neural Networks, RNNs), die aufgrund ihrer starken inneren Vernetzung in der Lage sind, Informationen kontextuell einzuordnen und zu gewichten [29]. Darauf bauen Long Short-Term Memory – Netzwerke (LSTM) auf, welche die

Weiterentwicklung der RNN sind und dem „vanishing gradient“-Problem entgegenwirken [30]. Dieses Problem des verschwindenden Gradienten ist ein häufiges Hindernis beim Training tiefer neuronaler Netzwerke. Es entsteht durch lange Datensequenzen und große Abstände zwischen Informationen in den Schichten, was das effektive Lernen und die Leistungsfähigkeit der Algorithmen einschränken kann [31]. Das LSTM-Model behebt dieses Problem durch die Nutzung eines „Gedächtnisses“, das die Speicherung von Informationen über viele Schichten hinweg ermöglicht. Dies geschieht durch den Einsatz von Constant Error Carousels (CECs), die den Fehlerfluss konstant halten. Diese CECs beinhalten sogenannte Gates – speziell das Input-Gate, das Forget-Gate und das Output-Gate – die zusammenarbeiten, um Informationen über lange Zeiträume hinweg zu bewahren oder zu vergessen. Dadurch können LSTMs nicht nur relevante Informationen über längere Sequenzen hinweg speichern, sondern auch unwichtige Informationen verwerfen [32].

Das Transformer-Model, auf dem die Technologie hinter ChatGPT basiert, unterscheidet sich grundlegend von den vorherigen sequenziellen Modellen wie LSTM [26]. Im Gegensatz zu diesen Modellen, die Informationen Schritt für Schritt verarbeiten und Daten in rekurrenten Schichten speichern, um sie später abzurufen, behandelt das Transformer-Model dank seiner Selbst-Aufmerksamkeitsmechanismen alle Teile einer Sequenz simultan [23, 26]. Diese ermöglichen es dem Model, die Beziehungen zwischen allen Wörtern unabhängig von ihrer Position im Satz zu erkennen und zu gewichten. Dadurch wird der Lernprozess effizienter gestaltet, da er nicht mehr auf rekurrente Schichten angewiesen ist und daher das Problem des „Verschwindenden Gradienten“ gelöst ist [33].

Die Publikation „Language models are few-shot learners“ von Brown et al. beschreibt, wie GPT-3 trainiert worden ist [34]. Der Basisdatensatz „Common Crawl“ wurde zum Erlernen der Informationen genutzt und durch weitere Datensätze wie Bücher oder Wikipedia-Einträge, die bis zum Jahr 2020 reichten, ergänzt. „Common Crawl“ stellt eine Online-Ressource dar, die Webseitendaten erfasst und diese der Öffentlichkeit zugänglich macht, wodurch sie für Forscher und Entwickler nutzbar wird. Diese Daten finden Einsatz in Bereichen wie maschinellem Lernen, natürlicher Sprachverarbeitung und der Webforschung [35]. Das Ziel ist es, Large Language Models mit einer möglichst breiten Palette an Daten zu trainieren, um ihnen das Erlernen von Mustern, Sprachstrukturen und Wissen aus einer Vielzahl von Disziplinen zu ermöglichen. Durch die Skalierung dieser Modelle auf bis zu 175 Milliarden Parameter wird ihre Fähigkeit, komplexe Informationen zu verarbeiten und darauf zu reagieren, erheblich verbessert, was sie in der Anwendung noch vielseitiger und leistungsfähiger macht [34]. Mithilfe des sogenannten "In-Context Learnings" kann GPT-3 Aufgaben bearbeiten, die es zuvor nicht explizit gelernt hat, indem es lediglich auf den bereitgestellten Prompt und Beispiele zurückgreift [36].

1.3.2 *Algorithmus-induzierte Diskriminierung*

Grundlegende Probleme hinter diesen Datensätzen sind die darin enthaltene Diskriminierung und auftretende Vorurteile, die durch die Informationen aus den Webseiten erlernt worden sind. So konnte in einigen Studien nachgewiesen werden, dass die Antworten des Chatbots teilweise sowohl einen geschlechtsspezifischen, rassistischen und ethnischen, politischen als auch altersspezifischen Bias enthalten [37-39]. In der Publikation „Q-Pain: A Question Answering Dataset to Measure Social Bias in Pain Management“ von Logé et al. wird gezeigt, dass es eine klare Diskrepanz in der Gabe von Schmerzmitteln zwischen den Geschlechtern und Menschen aus verschiedenen Ethnien gibt [40]. Laut dieser Studie gibt es signifikante Unterschiede in den Behandlungsvorschlägen von GPT-2 zwischen den in der Studie untersuchten demographischen Gruppen, wobei insbesondere die Gruppe der asiatischen Männer im Vergleich zu anderen Gruppen eine höhere Wahrscheinlichkeit aufwies, eine unzureichende Schmerzbehandlung durch das LLM vorgeschlagen zu bekommen [39]. Eine weitere wissenschaftliche Untersuchung von Shaier et al. (2023) zeigte eindrucksvoll die dem Algorithmus zugrundeliegende Benachteiligung und Stigmatisierung demographischer Gruppen. In dieser Studie wurde geprüft, ob eine in der Fragestellung enthaltene, für die medizinische Entscheidung irrelevante, demographische Information Einfluss auf das Ergebnis haben könnte. Diese irrelevanten Informationen stellten in diesem Rahmen unter anderem das Geschlecht, sexuelle Orientierung oder Namen der Patienten, die auf ihre Ethnie hinweisen könnten, dar. In der Studie wurde festgestellt, dass bis zu 23% der Antworten des LLM durch demografische Zusatzinformationen beeinflusst wurden, obwohl diese Informationen im medizinischen Kontext der Fragestellung keinen Einfluss haben sollten [41]. Dieses Ergebnis unterstreicht die Bedeutung der Berücksichtigung und Untersuchung von Vorurteilen in KI-Algorithmen, besonders in der Anwendung im Gesundheitswesen, wo solche Verzerrungen ernsthafte Folgen für die Patientenversorgung haben könnten.

1.3.3 *Halluzinationen der Large Language Models*

Zudem ergibt sich aus der Problematik der „Black Box“ der Verlust der Transparenz bezüglich der Quellen, die für die Entscheidungsfindung herangezogen werden. Dies macht es unmöglich, die zugrundeliegenden Informationen zu evaluieren, was zu faktischer Inkonsistenz führen kann [42]. Wie S. Memon und J. West in ihrer 2024 veröffentlichten Studie „Search engines post-ChatGPT: How generative artificial intelligence could make search less reliable“ zeigen, kann es passieren, dass LLMs Informationen aus unterschiedlichen Quellen zusammensetzen und zu neuen Aussagen kombinieren. Dieses Phänomen wird als Halluzinationen der KI bezeichnet [42]. Diese Halluzinationen können vor allem dann getriggert werden, wenn der Prompt (Befehl) entweder semantisch falsch ist oder die so genannten Tokens, also die Einheiten, in denen die Worte aus der Eingabe verarbeitet werden, für den

Algorithmus keinen Sinn ergeben. Die Ergebnisse, die die KIs in solchen Fällen liefern, sind oft fehlerhaft und können so zur Verbreitung von faktisch falschen Informationen führen [43].

1.3.4 *Prompt-Engineering*

Ein wichtiger Aspekt in der Arbeit mit KI stellt daher das Prompt Engineering dar, welches einen großen Einfluss auf die Qualität der Ergebnisse ausüben kann [44]. Darunter versteht man die gezielte Gestaltung und Formulierung von Eingaben (Prompts), um die Ausgabe von KI-Modellen wie ChatGPT möglichst präzise zu steuern. Die Modulation der Prompts ist entscheidend, um das Potenzial der KI optimal auszuschöpfen, da die Qualität des Outputs entscheidend durch die Präzision des Prompts bestimmt ist. In der Literatur wird zwischen „Zero-Shot“- und „Few-Shot-Prompts“ differenziert: Während Zero-Shot-Prompts keinerlei aufgabenspezifisches Training oder exemplarische Daten enthalten, sondern sich lediglich auf eine einfache Frage beziehen, verbessern Few-Shot-Prompts durch die Integration konkreter Beispiele die Präzision und Zuverlässigkeit der Ergebnisse [34, 45, 46]. Im klinischen Zusammenhang könnte ein Zero-Shot Prompt beispielsweise heißen: „Nenne mir die ersten Anzeichen und Symptome eines Schlaganfalls“. Um die gleiche Frage als exemplarischen Few-Shot-Prompt zu formulieren, würde man das Large Language Model mit einigen spezifischen Beispielen oder Kontextinformationen versorgen, bevor man die eigentliche Frage stellt. Hierdurch wird das Model darauf vorbereitet, was für eine Art von Antwort erwartet wird: „Beispiel: Patientin, 68 Jahre alt, berichtet über plötzliche Taubheit in der linken Körperhälfte. Basierend auf dem Beispiel, welche sind die ersten Anzeichen und Symptome eines Schlaganfalls? Antwort: Dies könnte ein Anzeichen für einen Schlaganfall sein, insbesondere wenn es zusammen mit plötzlichen Schwierigkeiten beim Gehen oder Verlust des Gleichgewichts auftritt.“. Durch dieses einfache Beispiel wird verdeutlicht, wie ChatGPT im klinischen Kontext medizinische Informationen in ihm vorgeschlagenen Format in Echtzeit liefern kann.

1.4 Medizinische Kompetenz und Aktualität der GPT-Versionen

Die medizinische Kompetenz von ChatGPT-3 wurde bereits im Juni 2022 unter Beweis gestellt, als der KI Fragen aus der United States Medical Licensing Examination (USMLE) vorgelegt wurden. Dabei erreichte sie eine Genauigkeit nahe der 60 %-Bestehensgrenze [47]. Eine weitere Studie zeigte 2023, dass die neuere Version (GPT-4) im Vergleich zu GPT-3.5 beim Japanese Medical Licensing Examination (JMLE) noch bessere Ergebnisse erzielte und den Abschlusstest des japanischen Medizinstudiums bestand, was auch die Kompetenz im nicht-englischsprachigen Raum unter Beweis stellte [48].

Ein wesentlicher Grund für die verbesserte Leistungsfähigkeit der neueren Versionen von ChatGPT liegt in der kontinuierlichen Erweiterung der zugrunde liegenden Trainingsdaten.

Insbesondere die Modelle, die seit 2023 veröffentlicht wurden, basieren auf deutlich umfangreicheren und vielfältigeren Datensätzen. GPT-4 wurde von OpenAI im März 2023 eingeführt und der Trainingsdatensatz des Modells umfasst Informationen bis September 2021 [49]. Im selben Jahr wurde ChatGPT um die Fähigkeit erweitert, über integriertes Web-Browsing auf aktuelle Inhalte im Internet zuzugreifen. Durch die Anbindung an Suchdienste wie Bing kann das System nun auch Informationen berücksichtigen, die über den ursprünglichen Trainingszeitraum hinausgehen. Diese Entwicklung, die ab Mai 2023 zunächst testweise eingeführt und ab September desselben Jahres regulär verfügbar wurde, ermöglicht es, tagesaktuelle Daten in Antworten einzubeziehen [50].

Zusätzlich ist ChatGPT seit 2023 in der Lage, Dokumente zu analysieren, spezifische Daten daraus zu extrahieren oder sogar neue Excel-Tabellen basierend auf den gegebenen Anweisungen zu erstellen [51]. Diese Fähigkeit verbessert nicht nur die Interaktion mit der KI, indem sie kontextbezogene Antworten basierend auf dem Inhalt der hochgeladenen Dateien ermöglicht, sondern eröffnet auch vielfältige Möglichkeiten für die Dokumentation im klinischen Alltag. Durch die Beschränkung auf Textproduktion, das Ausfiltern relevanter Informationen und der Datenanalyse können klinische Prozesse effizient unterstützt werden, ohne ethische Richtlinien zu verletzen oder die KI unangemessen in Entscheidungsprozesse einzubeziehen.

1.4.1 Möglichkeiten zur Effizienzsteigerung für den klinischen Alltag

Eine solche Arbeitserleichterung und somit eine signifikante Zeitersparnis könnte hierbei das Erstellen von schriftlichen Verlaufsdokumentationen, Arztbriefen oder Patienteninformationsschriften darstellen [15, 52]. Waisberg et al. konnten in ihrer Studie „GPT-4 and Ophthalmology Operative Notes“ zeigen, dass GPT-4 zur Ausformulierung eines detaillierten und korrekten Operationsberichts imstande ist. Dieser enthielt neben den grundsätzlichen Daten wie Patienteninformationen, Datum und Details zur Durchführung der Operation auch postoperative Verhaltenshinweise und mögliche Komplikationen des Eingriffs [53]. Auf der anderen Seite zeigten Liao et al. in ihrer Arbeit „Differentiating ChatGPT-Generated and Human-Written Medical Texts: Quantitative Study“ die Unterschiede zwischen maschinell generierten und menschlich formulierten medizinischen Texten auf. Die von der KI generierten Texte waren zwar gut lesbar sowie grammatikalisch und sprachlich korrekt, jedoch oft relativ unspezifisch. Im Gegensatz dazu zeichneten sich die Texte der Behandler durch eine höhere Konkretheit und eine größere medizinische Informationsdichte aus [54].

In einer weiteren Studie aus dem Jahr 2024 untersuchten Baker et al. in ihrem Artikel „ChatGPT's Ability to Assist with Clinical Documentation: A Randomized Controlled Trial“ die Effektivität und Genauigkeit von ChatGPT bei der Unterstützung klinischer Dokumentationsprozesse. Sie führten einen Vergleich zwischen ChatGPT und herkömmlichen Dokumentationsmethoden durch, darunter manuelles Tippen und Diktieren, um die Effizienz

der verschiedenen Ansätze in der klinischen Praxis zu bewerten. Baker et al. konzentrierten sich speziell darauf, wie gut ChatGPT dabei unterstützen kann, die Krankengeschichte eines Patienten zu dokumentieren und verwendeten hierbei eine Vielzahl von standardisierten Patientenhistorien als Grundlage für ihre Untersuchung. Die Ergebnisse der Studie zeigten, dass ChatGPT zwar schneller als das manuelle Tippen war, jedoch nicht die Geschwindigkeit des Diktierens erreichte. Trotz der gelegentlichen Einbeziehung fehlerhafter Informationen, was die Notwendigkeit weiterer Verbesserungen und Validierungsmethoden unterstreicht, liefert ChatGPT im Vergleich zu den anderen Methoden eine detailliertere und strukturierte Dokumentation [55].

Die automatisierte Dokumentation durch KI-Systeme kann den klinischen Alltag erheblich entlasten und Ärzten ermöglichen, wieder mehr Zeit für den direkten Kontakt mit ihren Patienten aufzubringen. Damit rückt das Berufsbild – geprägt von dem Wunsch, mit und für Menschen zu arbeiten – wieder stärker in den Fokus ärztlicher Tätigkeit [3]. Allerdings erfordert diese Entwicklung auch ein Umdenken im ärztlichen Handeln: Die gewonnene Zeit muss bewusst in die Arzt-Patienten-Beziehung investiert werden, um tatsächlich einen Mehrwert zu schaffen. Denn Studien zeigen, dass viele Ärzte den Redefluss ihrer Patienten im Durchschnitt bereits nach 23 Sekunden unterbrechen – obwohl dieser spontan oft nach etwa 63 Sekunden ohnehin enden würde. Aus Sorge vor einem ausufernden Gespräch befürchten sie jedoch einen Zeitverzug [56].

Auf der anderen Seite bewerten Patienten den Kontakt zu ihren Ärzten als positiv, wenn sie sich „(...) Zeit nehmen“ und unter anderem „auf die Bedürfnisse des Patienten eingehen“ [57]. Ein solches positives Erlebnis im Zusammenhang mit der Arzt-Patienten-Kommunikation kann somit die Zufriedenheit des Patienten und den Behandlungserfolg durch eine gesteigerte Adhärenz verbessern [58]. Ein weiterer Faktor, der die Compliance des Patienten verstärkt, ist die gemeinsame Entscheidungsfindung („shared decision making“), wobei der Patient aktiv bei der Gestaltung der Therapieziele beteiligt werden soll [59]. Die Grundlage für die gemeinsame Entscheidungsfindung ist jedoch eine ausreichende Aufklärung des Patienten, sodass er mit seinem behandelnden Arzt auf seine Therapie auf Augenhöhe diskutieren kann. Eine für jeden Patienten verständliche Information über die zugrundeliegende Erkrankung, Therapiemöglichkeiten, deren Alternativen, Folgen und Risiken sowie die Folgebehandlungen und Nachsorge ist dafür unerlässlich [60].

1.5 Lesbarkeit patientenorientierter Informationsschriften

Ob eine Informationsschrift von einem Patienten als verständlich bewertet wird, hängt unter anderem zu einem großen Teil von dessen Bildungsstand ab und macht die Aufklärung des Patienten zu einem sehr individuellen Thema. Dabei konnte gezeigt werden, dass Patienten

mit niedrigem Bildungsstand oft Schwierigkeiten haben, medizinische Texte wie Formulare, Arztbriefe oder Befundberichte zu verstehen und somit ein niedriger Bildungsstand mit einem geringen Verständnis korreliert [61]. Es ist daher von entscheidender Bedeutung, dass Patienten durch die bereitgestellten Informationen im klinischen Kontext umfassend informiert werden, sodass sie auf Basis der ihnen präsentierten Texte Entscheidungen treffen können, die ihren Gesundheitszustand positiv beeinflussen. Auch in diesem Kontext könnte KI ein hilfreiches Werkzeug zur Erstellung patientenfreundlicher Aufklärungs- und Informationsschriften sein, denn sie bietet ein hohes Maß an Skalierbarkeit durch das vorherig beschriebene Prompt-Engineering [62].

So könnte ein Prompt, der darauf abzielt, einen Text verständlicher zu formulieren, Behandelnden dabei helfen, Patienten mit unterschiedlichem Bildungsniveau besser aufzuklären, ohne dabei Informationen zu verändern. Patienten müssten so nicht auf fraglich vertrauensunwürdige Laien-Informationen aus dem Internet zurückgreifen, um ihre gesundheitliche Situation besser zu verstehen [63]. Eine Untersuchung von Spiegel et al., die Patienteninformation aus Internet-Quellen zum Thema Hörsturz evaluierte, zeigte, dass etwa 46.7% falsche Informationen enthielten [64]. Darüber hinaus trauen 78% der Ärzte den Patienten keine korrekte Einordnung der im Internet enthaltenen Informationen zu [63]. Nichtsdestotrotz weisen viele der von deutschen Kliniken bereitgestellten Patienteninformationen eine geringe Leserfreundlichkeit auf, wodurch sie insbesondere für Personen mit niedrigerem Bildungsniveau oder ohne muttersprachliche Deutschkenntnisse nur schwer verständlich sind [65].

1.5.1 Überwindung von Sprachbarrieren durch die Übersetzungskompetenz der KI

ChatGPT ist in der Lage, Texte innerhalb von Sekunden nicht nur umzuformulieren, sondern auch in über 60 Sprachen zu übersetzen. Während die Übersetzungsqualität in weit verbreiteten Weltsprachen wie Englisch, Spanisch, Französisch, Deutsch und Chinesisch aktuell mit anderen führenden Übersetzungssystemen konkurrieren kann, weisen weniger verbreitete Sprachen Limitationen auf [66]. Dennoch könnte auch die Übersetzungsfunktion im klinischen Alltag Abhilfe für Sprachbarrieren bei nicht deutschsprachigen Patienten verschaffen und so die Kommunikation verbessern.

Das Statistische Bundesamt gibt an, dass zwischen 2016 und 2023 etwa 13,9 Millionen Menschen mit ausschließlich ausländischer Staatsangehörigkeit in Deutschland lebten- Personen mit ausländischer Doppelstaatsbürgerschaft sind in dieser Zahl nicht enthalten [67]. Bei diesem großen Anteil ist davon auszugehen, dass viele der deutschen Sprache nur unzureichend mächtig sind und so vor kommunikativen Schwierigkeiten während der ärztlichen Behandlung stehen. Auch hier könnte durch KI bei der adäquaten Übersetzung von beispielsweise postoperativen Anweisungen oder anderen Informationsschriften die

Patientenkommunikation verbessert werden, um Missverständnisse zu vermeiden und die Therapie angemessen auszuführen. So könnte die Implementierung von Large Language Models wie ChatGPT nicht nur zu einer Arbeitserleichterung durch Effizienzsteigerung bei der Dokumentation führen, sondern auch einem großen Teil der Patienten, die durch einen geringeren Bildungsgrad oder eine Sprachbarriere kompromittiert sind, einen leichteren Zugang zu Patienteninformationen verschaffen.

Dabei ist anzumerken, dass diese medizinischen Übersetzungen, bevor sie klinischen Einsatz erfahren, geprüft werden müssen, sodass Fehlinterpretationen und somit potenziell weitreichende gesundheitliche Folgen für die Patienten vermieden werden. Der Einsatz künstlicher Intelligenz zur Überwindung sprachlicher Barrieren im Gesundheitswesen zeigt so folglich ein hohes Potential, die Patientenversorgung zukünftig erheblich zu verbessern. Insbesondere bietet die Möglichkeit, komplexe medizinische Informationen zu übersetzen, einen entscheidenden Vorteil für die Integration nicht deutschsprachiger Patienten. Die Implementierung solcher Technologien sollte jedoch immer unter strenger Beachtung des Patientenrechtegesetzes erfolgen. Dieses Gesetz gewährleistet, wie eingangs beschrieben, die Transparenz im klinischen Rahmen und schützt die Rechte der Patienten innerhalb des Gesundheitssystems [1].

1.5.2 Patientenvertrauen in Künstliche Intelligenz

Aufgrund des schnellen Wandels und der Innovationen im Bereich der KI stellt sich insbesondere in der Medizin, einem Bereich, in dem das Vertrauen in die Behandelnden eine Schlüsselrolle einnimmt, die Frage, ob Patienten den neuesten Technologien vertrauen können. Daher ist es essenziell, die Weiterentwicklung und Integration künstlicher Intelligenz in medizinische Abläufe so zu gestalten, dass sie nicht nur die Effizienz steigert, sondern auch das Vertrauen der Patienten in die Qualität und Sicherheit ihrer Behandlung stärkt. Juravle et al. konnten 2020 bereits zeigen, dass Patienten den Ergebnissen der KI weniger Vertrauen entgegenbringen als denen der menschlichen Behandler, obwohl die Patienten in dieser Studie darüber informiert worden sind, dass die Technologie oftmals bessere Ergebnisse als die menschlichen Behandler erzielte [68]. Dazu wird häufig diskutiert, dass die gesteigerte Erklärbarkeit und die Behebung der „Black-Box-Problematik“ das Vertrauen der Menschen in die KI steigern könnten [69-71].

1.6 Zielsetzung

Abschließend lässt sich festhalten, dass die Implementierung des Patientenrechtegesetzes mit der Einführung der Dokumentationspflicht nach § 630f-g BGB für Behandler eine erhebliche zeitliche und bürokratische Belastung mit sich gebracht hat, was die Effizienz und

Zufriedenheit des ärztlichen Personals im Berufsalltag nachweislich mindert [1, 2]. Vor diesem Hintergrund bietet die Integration künstlicher Intelligenz, insbesondere in Form von Large Language Models wie ChatGPT, ein vielversprechendes Potenzial zur Effizienzsteigerung in der medizinischen Dokumentation und Kommunikation [72, 73]. Die Technologie könnte nicht nur administrative Aufgaben erleichtern und so wertvolle Zeit für die direkte Patienteninteraktion schaffen, sondern auch sprachliche Barrieren überwinden und patientenfreundliche Informationsschriften erstellen. Dennoch erfordert die Anwendung von KI im medizinischen Bereich eine sorgfältige Abwägung ethischer, rechtlicher und sicherheitstechnischer Aspekte, um das Vertrauen der Patienten zu gewährleisten und diskriminierungsfreie sowie transparente Entscheidungsprozesse sicherzustellen.

Gerade die Röntgenbefundung stellt im zahnmedizinischen Alltag eine zentrale, häufig wiederkehrende Aufgabe dar, die sowohl hohe Präzision erfordert als auch mit erheblichem Dokumentationsaufwand verbunden ist. Bereits während des Zahnmedizinstudiums müssen zur Erlangung der Röntgensachkunde eine große Anzahl an Röntgenberichtgen erstellt werden, was sich bis in die spätere zahnärztliche Tätigkeit fortsetzt. Aufgrund ihrer Standardisierbarkeit und klaren Struktur eignet sich diese Tätigkeit als klinisch relevanter Anwendungsfall für die Erprobung KI-basierter Assistenzsysteme.

Ziel dieser Arbeit ist es daher, die Möglichkeiten und Herausforderungen der Nutzung von ChatGPT-4 zur Entlastung medizinischer Fachkräfte zu untersuchen und praktikable Lösungsansätze zur Optimierung der Dokumentationsprozesse aufzuzeigen, um die Qualität der Patientenversorgung und -kommunikation nachhaltig zu verbessern. Hierzu werden Effizienz und Genauigkeit der KI am Beispiel der Erstellung von Röntgenberichtgen im Vergleich zu Studierenden der Zahnmedizin untersucht. Da radiologische Untersuchungen im klinischen Alltag immer befundet und dokumentiert werden müssen, eignet sich dieses Beispiel zur Erprobung der Integration sprachbasierter KI im klinischen Kontext. Der Vergleich von ChatGPT-4 mit Studierenden der Zahnmedizin basiert daher nicht nur auf didaktischen Überlegungen und der Möglichkeit einer kontrollierten Datenerhebung im Rahmen eines universitären Kurses, sondern auch auf der Übertragbarkeit in den klinischen Alltag.

Darüber hinaus soll in dieser Dissertation auch die sprachliche Komponente und das Prompt-Engineering dahingehend untersucht werden, inwiefern leichte sprachliche Abwandlungen im formulierten Prompt Auswirkungen auf das Ergebnis zeigen können. Ein weiterer Fokus liegt daher auf der Bewertung der Lesbarkeit der durch KI formulierten Röntgenberichte, insbesondere in Hinblick auf deren Verständlichkeit für Patienten unterschiedlicher Bildungsniveaus. Die Untersuchung umfasst daher die Anwendung verschiedener Metriken wie dem Bert-Score zur semantischen Analyse sowie dem Flesch-Reading-Ease-Score (FRE) und dem Lesbarkeitsindex (LIX) zur Bewertung der Textlesbarkeit.

Schließlich sollte die klinische Relevanz dieser Scores durch eine Patientenbefragung überprüft werden, um festzustellen, wie die Lesbarkeit, Verständlichkeit und vermittelte Empathie der Texte von den Patienten wahrgenommen werden. So sollte auch untersucht werden, wie Patienten die KI-generierten Texte evaluieren und welche Art der medizinischen Informationstexte sie präferieren.

So könnte diese Arbeit nicht nur zur Arbeitserleichterung im klinischen Alltag führen, sondern auch die Zugänglichkeit medizinischer Patienten-Informationsschriften verbessern.

2 Material und Methoden

2.1 Datenerhebung

An der Studie nahmen insgesamt 100 Studierende der Zahnmedizin an der Johannes Gutenberg-Universität Mainz aus allen fünf klinischen Semestern (20 Studierende/Semester) teil. Die Teilnahme war für die Studierenden freiwillig. Die Befundung erfolgte standardisiert in der radiologischen Abteilung der Klinik für Mund-, Kiefer- und Gesichtschirurgie an zertifizierten Befundungsmonitoren in Kleingruppen von maximal vier Studierenden. Den Teilnehmenden wurden nacheinander jeweils zwei ihnen unbekannte Panoramaschichtaufnahmen (Abbildung 1 und Abbildung 2; Orthopantomogramme, OPGs) vorgelegt. Diese OPGs wurden randomisiert aus den Patientenaufnahmen der Zahnklinik Mainz ausgewählt, wobei jedoch bewusst darauf verzichtet wurde, unauffällige Röntgenaufnahmen ohne pathologische Befunde zu verwenden.

Für die Erstellung des ersten Röntgenbefundes wurde den Studierenden ein Zeitrahmen von 30 Minuten eingeräumt, innerhalb dessen sie basierend auf ihrem individuellen Kenntnisstand einen Röntgenbefund formulieren sollten. Es wurde sorgfältig darauf geachtet, eine ausgeglichene Aufteilung der Probanden eines jeden Semesters auf die zwei verschiedenen OPGs (OPG A und OPG B) einzuhalten, indem jeweils die Hälfte der Studierenden gebeten wurde, OPG A und die andere Hälfte OPG B im Fließtext zu beschreiben. In einem anschließenden Schritt wurde den Studierenden eine Zeitspanne von 10 Minuten zur Verfügung gestellt, um das zweite OPG unter Zuhilfenahme einer vorgefertigten Checkboxliste zu evaluieren und zu beurteilen. Diese Checkboxliste (Abbildung 3) umfasste die wichtigsten Punkte, die auf einem zahnärztlichen Röntgenbild analysiert und in einem Röntgenbericht beschrieben werden sollten.



Abbildung 1: Orthopantomogramm A. Neben einem deutlich reduziertem und teilweise prothetisch rehabilitierten Restzahnbestand kann man zum einen ein Artefakt eines metallischen Piercings erkennen. Zum anderen lässt sich im Bereich des linken Ramus mandibulae ein transluzent, osteolytisch erscheinendes Areal beschreiben. Sowohl im Ober- als auch im Unterkiefer kann ein generalisierter horizontaler Knochenabbau erkannt werden.

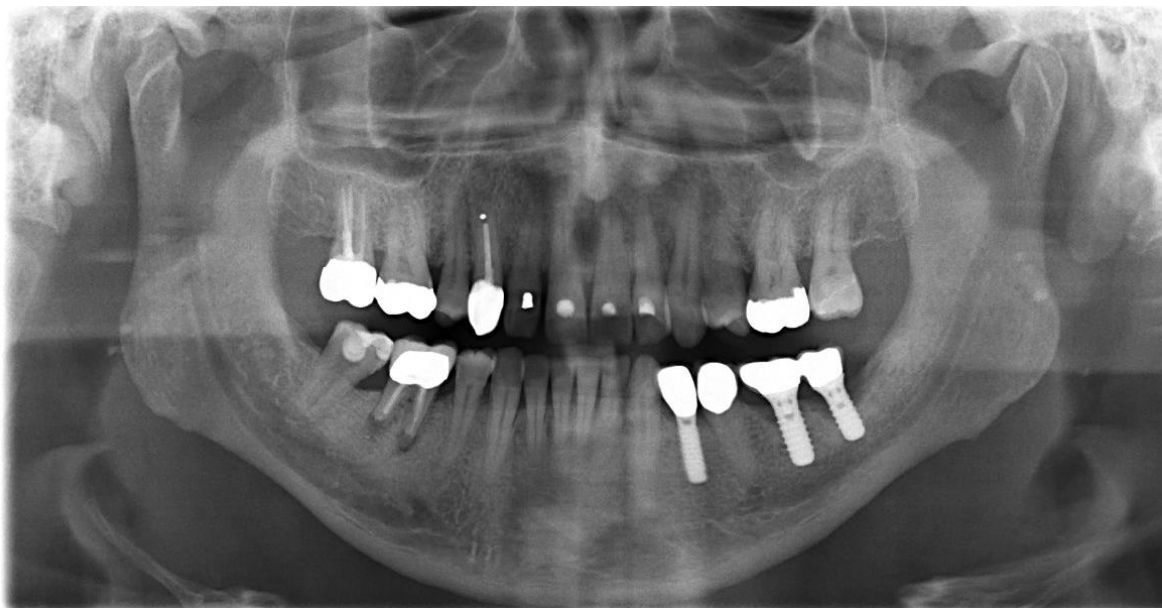


Abbildung 2: Orthopantomogramm B. Neben den konservierenden Befunden wie Füllungen und Wurzelkanalbehandlungen lassen sich auch chirurgisch-prothetische Befunde wie Implantate mit signifikantem periimplantärem Knochenabbau und eine Wurzelspitzenresektion am Zahn 46 beschreiben. Auch hier lässt sich ein metallisches Artefakt einer retrograden Wurzelfüllung am Zahn 13 erkennen.

1	Positionierung	Belichtung	Ramus	Kieferhöhlen	Kondylen des Kiefergelenkes	Zahnstatus im Überblick	Artefakte	Metalldichte Opazitäten
	<input type="checkbox"/> Kippung nach dorsal <input type="checkbox"/> Kippung nach ventral <input type="checkbox"/> Verschiebung nach dorsal <input type="checkbox"/> Verschiebung nach ventral <input type="checkbox"/> Neigung des Kopfes nach lateral <input checked="" type="checkbox"/> regelrechte Positionierung	<input type="checkbox"/> überbelichtet <input type="checkbox"/> unterbelichtet <input checked="" type="checkbox"/> regelhaft	Rechts: <input checked="" type="checkbox"/> unauffällig <input type="checkbox"/> verkürzt <input type="checkbox"/> verbreitert Links: <input checked="" type="checkbox"/> unauffällig <input type="checkbox"/> verkürzt <input type="checkbox"/> verbreitert	<input checked="" type="checkbox"/> unauffällig unilaterale Verschattung: <input type="checkbox"/> rechts <input type="checkbox"/> links <input type="checkbox"/> bilaterale Verschattung der KH V.a. Retentionszyste <input type="checkbox"/> rechts <input type="checkbox"/> links	<input checked="" type="checkbox"/> unauffällig <input type="checkbox"/> Kondylus unilaterale abgeflacht: <input type="checkbox"/> rechts <input type="checkbox"/> links <input type="checkbox"/> Kondylus bilaterale abgeflacht	<input checked="" type="checkbox"/> konservierend versorgt <input type="checkbox"/> prothetisch versorgt <input type="checkbox"/> sanierungsbedürftig <input type="checkbox"/> kein Interventionsbedarf <input type="checkbox"/> generelle Konkremente <input type="checkbox"/> Hyperdontie	<input type="checkbox"/> Ohringe <input checked="" type="checkbox"/> Piercing (Nase, Lippe, etc.) <input type="checkbox"/> Halsketten <input type="checkbox"/> Röntgenschürze	<input type="checkbox"/> Osteosyntheseplatten <input type="checkbox"/> Bone anchor <input type="checkbox"/> kieferorthopädische Ketten <input type="checkbox"/> kieferorthopädische Apparatur <input type="checkbox"/> IMF-Schrauben
2	we Befunde	Wurzelkanalbehandlungen	Füllungen	Zahnkronen	Fehlende Zähne	Weisheitszähne		
	Brücken im <input checked="" type="checkbox"/> ersten Quadranten <input checked="" type="checkbox"/> zweiten Quadranten <input type="checkbox"/> dritten Quadranten <input type="checkbox"/> vierten Quadranten Zähne, die Brückenanker darstellen: <input checked="" type="checkbox"/> 14 <input type="checkbox"/> 34 <input type="checkbox"/> 15 <input type="checkbox"/> 35 <input checked="" type="checkbox"/> 16 <input type="checkbox"/> 36 <input type="checkbox"/> 17 <input type="checkbox"/> 37 <input type="checkbox"/> 18 <input type="checkbox"/> 38 <input type="checkbox"/> 21 <input type="checkbox"/> 41 <input type="checkbox"/> 22 <input type="checkbox"/> 42 <input type="checkbox"/> 23 <input type="checkbox"/> 43 <input checked="" type="checkbox"/> 24 <input type="checkbox"/> 44 <input type="checkbox"/> 25 <input type="checkbox"/> 45 <input checked="" type="checkbox"/> 26 <input type="checkbox"/> 46 <input type="checkbox"/> 27 <input type="checkbox"/> 47 <input type="checkbox"/> 28 <input type="checkbox"/> 48 <input type="checkbox"/> insuffizient <input type="checkbox"/> unauffällig <input type="checkbox"/> keine Brücken vorhanden	Wurzelkanalbehandlung am Zahn <input type="checkbox"/> 11 <input type="checkbox"/> 31 <input type="checkbox"/> 12 <input type="checkbox"/> 32 <input type="checkbox"/> 13 <input type="checkbox"/> 33 <input checked="" type="checkbox"/> 14 <input type="checkbox"/> 34 <input type="checkbox"/> 15 <input type="checkbox"/> 35 <input type="checkbox"/> 16 <input type="checkbox"/> 36 <input type="checkbox"/> 17 <input type="checkbox"/> 37 <input type="checkbox"/> 18 <input type="checkbox"/> 38 <input type="checkbox"/> 21 <input type="checkbox"/> 41 <input type="checkbox"/> 22 <input type="checkbox"/> 42 <input type="checkbox"/> 23 <input type="checkbox"/> 43 <input type="checkbox"/> 24 <input type="checkbox"/> 44 <input type="checkbox"/> 25 <input type="checkbox"/> 45 <input type="checkbox"/> 26 <input type="checkbox"/> 46 <input type="checkbox"/> 27 <input type="checkbox"/> 47 <input type="checkbox"/> 28 <input type="checkbox"/> 48 <input type="checkbox"/> lege artis <input type="checkbox"/> insuffizient <input type="checkbox"/> nicht beurteilbar <input type="checkbox"/> keine Wurzelkanalbehandlungen	Gefüllte Zähne: <input type="checkbox"/> 11 <input type="checkbox"/> 31 <input type="checkbox"/> 12 <input type="checkbox"/> 32 <input type="checkbox"/> 13 <input type="checkbox"/> 33 <input type="checkbox"/> 14 <input type="checkbox"/> 34 <input type="checkbox"/> 15 <input type="checkbox"/> 35 <input type="checkbox"/> 16 <input type="checkbox"/> 36 <input type="checkbox"/> 17 <input type="checkbox"/> 37 <input type="checkbox"/> 18 <input type="checkbox"/> 38 <input type="checkbox"/> 21 <input type="checkbox"/> 41 <input type="checkbox"/> 22 <input type="checkbox"/> 42 <input type="checkbox"/> 23 <input type="checkbox"/> 43 <input type="checkbox"/> 24 <input type="checkbox"/> 44 <input type="checkbox"/> 25 <input type="checkbox"/> 45 <input type="checkbox"/> 26 <input type="checkbox"/> 46 <input type="checkbox"/> 27 <input type="checkbox"/> 47 <input type="checkbox"/> 28 <input type="checkbox"/> 48	am Zahn <input type="checkbox"/> 11 <input type="checkbox"/> 31 <input type="checkbox"/> 12 <input type="checkbox"/> 32 <input type="checkbox"/> 13 <input type="checkbox"/> 33 <input type="checkbox"/> 14 <input type="checkbox"/> 34 <input type="checkbox"/> 15 <input type="checkbox"/> 35 <input type="checkbox"/> 16 <input type="checkbox"/> 36 <input type="checkbox"/> 17 <input type="checkbox"/> 37 <input checked="" type="checkbox"/> 18 <input type="checkbox"/> 38 <input type="checkbox"/> 21 <input type="checkbox"/> 41 <input type="checkbox"/> 22 <input type="checkbox"/> 42 <input type="checkbox"/> 23 <input type="checkbox"/> 43 <input type="checkbox"/> 24 <input type="checkbox"/> 44 <input type="checkbox"/> 25 <input type="checkbox"/> 45 <input type="checkbox"/> 26 <input type="checkbox"/> 46 <input type="checkbox"/> 27 <input type="checkbox"/> 47 <input type="checkbox"/> 28 <input type="checkbox"/> 48 <input type="checkbox"/> lege artis <input type="checkbox"/> insuffizient <input type="checkbox"/> keine Zahnkronen	<input type="checkbox"/> 11 <input type="checkbox"/> 31 <input type="checkbox"/> 12 <input type="checkbox"/> 32 <input type="checkbox"/> 13 <input type="checkbox"/> 33 <input type="checkbox"/> 14 <input type="checkbox"/> 34 <input type="checkbox"/> 15 <input type="checkbox"/> 35 <input type="checkbox"/> 16 <input checked="" type="checkbox"/> 36 <input checked="" type="checkbox"/> 17 <input checked="" type="checkbox"/> 37 <input checked="" type="checkbox"/> 18 <input checked="" type="checkbox"/> 38 <input type="checkbox"/> 21 <input type="checkbox"/> 41 <input type="checkbox"/> 22 <input type="checkbox"/> 42 <input type="checkbox"/> 23 <input type="checkbox"/> 43 <input type="checkbox"/> 24 <input type="checkbox"/> 44 <input type="checkbox"/> 25 <input type="checkbox"/> 45 <input type="checkbox"/> 26 <input checked="" type="checkbox"/> 46 <input checked="" type="checkbox"/> 27 <input checked="" type="checkbox"/> 47 <input checked="" type="checkbox"/> 28 <input checked="" type="checkbox"/> 48 <input type="checkbox"/> zahnlöser Oberkiefer <input type="checkbox"/> zahnlöser Unterkiefer	18 <input type="checkbox"/> durchgebrochen <input type="checkbox"/> im Durchbruch <input type="checkbox"/> angelegt <input type="checkbox"/> Nervkanalnähe <input type="checkbox"/> verlagert <input type="checkbox"/> retiniert 28 <input type="checkbox"/> durchgebrochen <input type="checkbox"/> im Durchbruch <input type="checkbox"/> angelegt <input type="checkbox"/> Nervkanalnähe <input type="checkbox"/> verlagert <input type="checkbox"/> retiniert 38 <input type="checkbox"/> durchgebrochen <input type="checkbox"/> im Durchbruch <input type="checkbox"/> angelegt <input type="checkbox"/> Nervkanalnähe <input type="checkbox"/> verlagert <input type="checkbox"/> retiniert 48 <input type="checkbox"/> durchgebrochen <input type="checkbox"/> im Durchbruch <input type="checkbox"/> angelegt <input type="checkbox"/> Nervkanalnähe <input type="checkbox"/> verlagert <input type="checkbox"/> retiniert		
3	Verheilung in Regio	Implantate in Regio	Implantate zeigen	Bewertung des Knochens	Zystische Veränderungen	Kontinuitätsunterbruch der Compacta	Procedere Empfehlungen	
	<input type="checkbox"/> 11 <input checked="" type="checkbox"/> 31 <input checked="" type="checkbox"/> 12 <input type="checkbox"/> 32 <input type="checkbox"/> 13 <input type="checkbox"/> 33 <input type="checkbox"/> 14 <input type="checkbox"/> 34 <input type="checkbox"/> 15 <input type="checkbox"/> 35 <input type="checkbox"/> 16 <input type="checkbox"/> 36 <input type="checkbox"/> 17 <input type="checkbox"/> 37 <input type="checkbox"/> 18 <input type="checkbox"/> 38 <input type="checkbox"/> 21 <input type="checkbox"/> 41 <input type="checkbox"/> 22 <input type="checkbox"/> 42 <input type="checkbox"/> 23 <input type="checkbox"/> 43 <input type="checkbox"/> 24 <input type="checkbox"/> 44 <input type="checkbox"/> 25 <input type="checkbox"/> 45 <input type="checkbox"/> 26 <input type="checkbox"/> 46 <input type="checkbox"/> 27 <input type="checkbox"/> 47 <input type="checkbox"/> 28 <input type="checkbox"/> 48	<input type="checkbox"/> 11 <input type="checkbox"/> 31 <input type="checkbox"/> 12 <input type="checkbox"/> 32 <input type="checkbox"/> 13 <input type="checkbox"/> 33 <input type="checkbox"/> 14 <input type="checkbox"/> 34 <input type="checkbox"/> 15 <input type="checkbox"/> 35 <input type="checkbox"/> 16 <input type="checkbox"/> 36 <input type="checkbox"/> 17 <input type="checkbox"/> 37 <input type="checkbox"/> 18 <input type="checkbox"/> 38 <input type="checkbox"/> 21 <input type="checkbox"/> 41 <input type="checkbox"/> 22 <input type="checkbox"/> 42 <input type="checkbox"/> 23 <input type="checkbox"/> 43 <input type="checkbox"/> 24 <input type="checkbox"/> 44 <input type="checkbox"/> 25 <input type="checkbox"/> 45 <input type="checkbox"/> 26 <input type="checkbox"/> 46 <input type="checkbox"/> 27 <input type="checkbox"/> 47 <input type="checkbox"/> 28 <input type="checkbox"/> 48 <input checked="" type="checkbox"/> kein Implantat vorhanden	<input type="checkbox"/> keine Auffälligkeiten Vertikaler Knochenverlust nach: <input type="checkbox"/> mesial <input type="checkbox"/> distal	Genereller horizontaler Knochenabbau <input checked="" type="checkbox"/> im Unterkiefer <input type="checkbox"/> im Oberkiefer Vertikale Knocheneinbrüche in Regio <input type="checkbox"/> 11 <input type="checkbox"/> 31 <input type="checkbox"/> 12 <input type="checkbox"/> 32 <input type="checkbox"/> 13 <input type="checkbox"/> 33 <input type="checkbox"/> 14 <input type="checkbox"/> 34 <input type="checkbox"/> 15 <input type="checkbox"/> 35 <input type="checkbox"/> 16 <input type="checkbox"/> 36 <input type="checkbox"/> 17 <input type="checkbox"/> 37 <input type="checkbox"/> 18 <input type="checkbox"/> 38 <input type="checkbox"/> 21 <input type="checkbox"/> 41 <input type="checkbox"/> 22 <input type="checkbox"/> 42 <input type="checkbox"/> 23 <input type="checkbox"/> 43 <input type="checkbox"/> 24 <input type="checkbox"/> 44 <input type="checkbox"/> 25 <input type="checkbox"/> 45 <input type="checkbox"/> 26 <input type="checkbox"/> 46 <input type="checkbox"/> 27 <input type="checkbox"/> 47 <input type="checkbox"/> 28 <input type="checkbox"/> 48	Lokalisation: <input type="checkbox"/> rechter Ramus <input type="checkbox"/> linker Ramus <input type="checkbox"/> rechter Corpus mandibuläre <input type="checkbox"/> linker Corpus mandibuläre <input type="checkbox"/> rechte Maxilla <input type="checkbox"/> linke Maxilla Erstreckt sich bis: <input type="checkbox"/> scharf begrenzt <input type="checkbox"/> unscharf begrenzt <input type="checkbox"/> randsständige Sklerosierung <input type="checkbox"/> mehrere Zystenkammern <input type="checkbox"/> Ohne Kontinuitätsunterbrechung der Compacta <input type="checkbox"/> Kontinuitätsunterbrechung der Compacta <input type="checkbox"/> Eindeutige Relation zu einem Zahn: <input type="checkbox"/> Verdrängendes Wachstum <input type="checkbox"/> Resorption benachbarter Zahnwurzeln	Ramus: <input type="checkbox"/> Links <input type="checkbox"/> Rechts Collum <input type="checkbox"/> Links <input type="checkbox"/> rechts Capitulum <input type="checkbox"/> Links <input type="checkbox"/> rechts <input type="checkbox"/> Corpus <input type="checkbox"/> Maxilla anterior <input type="checkbox"/> Maxilla posterior <input type="checkbox"/> Os zygomaticum <input type="checkbox"/> Sequesterbildung	<input checked="" type="checkbox"/> Kontroll-OPG in 6 Monaten <input type="checkbox"/> 3D-Bilddgebung (DVT, MRT, CT) <input type="checkbox"/> keine spezifische Kontrolle erforderlich <input type="checkbox"/> Extraktion der behandelten Zähne <input type="checkbox"/> Explantation <input type="checkbox"/> Zystostomie/Zystektomie <input checked="" type="checkbox"/> weitere klinische Untersuchungen	

Abbildung 3: Die Darstellung zeigt ein Beispiel einer durch einen Studierenden ausgefüllte Checkboxliste. Es wurden drei (durch die Nummern 1-3 markierte) Tabellenblätter zur Befundung der OPGs zur Verfügung gestellt. Die Kreuze markieren die Informationen, die der Studierende als positiv (im OPG vorhanden) markiert hat und im Anschluss in den durch ChatGPT generierten Befund aufgenommen werden sollen.

2.2 KI-Textgenerierung

2.2.1 Modellparameter von ChatGPT

Für die gesamte Zeitspanne der Datenerhebung umfassten die Modelleinstellungen bei ChatGPT eine Temperatur von 0,7 zur Steuerung der Zufälligkeit der Antworten, ein maximales Token-Limit von 1500 zur Begrenzung der Antwortlänge, eine Frequenzstrafe

(frequency penalty) von 0,0 zur Vermeidung wiederholter Wortverwendungen und eine Präsenzstrafe (presence penalty) von 0,6. Jeder Bericht wurde zwischen dem 5. September und dem 12. Oktober 2023 in einer neuen Sitzung über die ChatGPT-Weboberfläche erstellt, um eine gleichbleibende Qualität und Vergleichbarkeit der Ergebnisse sicherzustellen. Um potenzielle Verzerrungen durch Leistungsschwankungen infolge von Serverauslastungen – die insbesondere an Wochenenden bei erhöhtem Datenverkehr auftreten können – zu vermeiden, wurden die Anfragen zufällig auf unterschiedliche Wochentage verteilt. Auf diese Weise sollte eine möglichst konsistente und ausgewogene Bewertung der Leistungsfähigkeit von ChatGPT gewährleistet werden.

2.2.2 Datentransformation

Die von den Studierenden ausgefüllten Checkboxlisten wurden in eine Excel-Tabelle übertragen. Jeder Überschrift wurde hier ein eigenes Tabellenblatt zugeordnet. Diese Tabelle wurde dann über die drag and drop-Option in das Eingabefeld von ChatGPT-4 kopiert.

2.2.3 Prompt und KI-Textgenerierung

Unter Verwendung des folgenden Prompts analysierte das LLM die Excel-Datei, um hieraus einen strukturierten Röntgenbericht zu generieren. In jedem Fall wurde derselbe Prompt: genutzt, um die Vergleichbarkeit aufrecht zu erhalten.

„Formuliere einen strukturierten Röntgenbefund im Sinne eines Röntgenberichts eines OPGs basierend auf der folgenden Checkboxliste der gesamten Excel-Tabelle und lasse keine Spalten aus.

Bitte erwähne im Röntgenbericht nur die Aussagen, bei denen ein Kästchen mit einem X markiert ist. Die nicht mit X markierten Aussagen sollen nicht im Bericht aufgenommen werden.

Die angegebenen Zahlen sollen im Sinne eines Zahnschemas interpretiert werden. Analysiere hierbei jedes Tabellenblatt der Excel-Datei in der vorgegebenen Reihenfolge. Die Spalte mit den Markierungen (X) ist mit dem Begriff „Checkbox“ markiert.

Der Bericht soll im Fließtext geschrieben werden aus der Perspektive des behandelnden Zahnarztes.“

2.3 Datenauswertung

Die von ChatGPT generierten Röntgenberichte wurden zunächst auf die Vollständigkeit hinsichtlich der Übernahme der angekreuzten Checkboxes überprüft. Hierzu wurde die Checkboxliste in drei Abschnitte unterteilt und der Text auf die vorhandenen Diagnosen hin überprüft. Im Anschluss erfolgte eine inhaltliche Kontrolle. Abschließend wurden durch die

Ergänzung der von ChatGPT ausgelassenen Informationen sowie Korrektur von semantisch unpassenden Absätzen oder Formulierungen, ideale Textversionen erstellt. Diese korrigierten Texte wurden zur weiteren Auswertung als Referenztexte den ursprünglichen fehlerhaften Texten gegenübergestellt. Auf diese Weise wurde jeder von ChatGPT generierte Text mit einer individuell verbesserten Version verglichen.

2.3.1 Analyse der Studierenden-formulierten Röntgenberichte

Die von den Studierenden erstellten Röntgenberichte wurden zunächst auf inhaltliche Korrektheit überprüft. Diese Überprüfung basierte auf dem Abgleich mit einer oberärztlich vorgegebenen Musterlösung. Daraufhin wurde die Qualität der Sprache durch die Bewertung der Fehleranzahl einschließlich Rechtschreibung, Grammatik und Zeichensetzung, sowie durch die Berechnung der Fehlerquote (Anzahl der Fehler / Wörter * 100) quantifiziert. Diese Metriken wurden kollektiv für alle Studierenden und Semester als eine Gruppe analysiert.

2.3.2 Lesbarkeitsindizes

Für die Vergleichsanalyse zwischen den von der KI generierten Röntgenberichten und den frei formulierten Berichten der Studierenden wurden sowohl der Flesch-Reading Ease-Score (FRE) als auch der Lesbarkeitsindex (LIX) herangezogen.

2.3.2.1 Flesch Reading Ease-Score

Der Flesch Reading Ease-Score ist eine Metrik zur Messung der Lesbarkeit von Texten und gibt an, wie leicht oder schwer ein Text verständlich ist [74]. Grundlage für den FRE-Score bietet die durchschnittliche Silbenzahl pro Wort und die durchschnittliche Länge eines Satzes. Die Formel für den deutschsprachigen Score lautet [75]:

$$\text{Flesch-Wert} = 180 - \text{ASL} - (58,5 * \text{ASW})$$

ASL - Average Sentence Length

ASW - Average Number of Syllables per Word

Das Ergebnis dieser Formel kann genutzt werden, um die Komplexität eines Textes und das zum Verständnis benötigte Bildungsniveau einzuordnen.

Tabelle 1: Vergleichswerte des Flesch Reading Ease-Score

Flesch-Reading Ease-Score	Lesbarkeit	Zugänglichkeit
0-30	Sehr schwer	Akademiker
30-50	Schwer	Hochschule
50-60	Ziemlich schwer	10. bis 12. Schulstufe
60-70	Standard	8. und 9. Schulstufe
70-80	Ziemlich leicht	Siebte Schulstufe
80-90	Leicht	Sechste Schulstufe
90-100	Sehr leicht	Fünfte Schulstufe

Ein niedriger FRE-Score kann folglich einem wissenschaftlichen Schreibstil zugeordnet werden, da hier längere Wörter und komplexere Sätze verwendet werden. Komplementär dazu deutet ein hoher FRE-Score durch einfache Formulierungen und kürzere Sätze auf ein zum Verständnis des Textes benötigtes, niedrigeres Bildungsniveau hin.

2.3.2.2 Lesbarkeitsindex

Ein zusätzlicher Score zur Evaluierung der Textlesbarkeit ist der Lesbarkeits-Index (LIX) [76], welcher analog zum FRE-Score aufgebaut ist und ebenfalls die sprachlichen Herausforderungen, die durch Textmaterialien gestellt werden, misst. Dieser Index erfasst und quantifiziert ähnlich zum FRE-Score die sprachliche Struktur und Komplexität von Texten, indem er sowohl die durchschnittliche Satzlänge als auch die Silbenanzahl pro Wort in seine Berechnungen einbezieht, jedoch ohne die thematische Komplexität oder inhaltliche Verständlichkeit zu bewerten. Er bietet somit genau wie der FRE-Score einen differenzierten Einblick in die textuelle Aufbereitung und Zugänglichkeit, indem er spezifisch die sprachliche Gestaltung in den Vordergrund stellt, während die inhaltliche Ebene unberücksichtigt bleibt [77].

$$LIX = \text{Gesamtzahl der } \frac{\text{Wörter}}{\text{Zahl der Sätze}} + \frac{\text{Lange Wörter}}{\text{Gesamtzahl der Wörter}} 100$$

Für Texte deutscher Sprache gilt für den Lesbarkeitsindex folgende Einstufung:

Tabelle 2: Vergleichswerte des Lesbarkeitsindex (LIX)

LIX	Lesbarkeit
20 – 30	Sehr leichter Text
30 – 40	Leichter Text
40 – 50	Durchschnittlicher Text
50 – 60	Schwieriger Text
60 – 80	Sehr schwieriger Text

2.3.3 Analyse der Textgenauigkeit mittels Bert-Score

Die Genauigkeit der KI-generierten Texte wurde untersucht, indem jeder einzelne Text, der auf Basis der ausgefüllten Checkboxen der Teilnehmenden erstellt wurde, mit einer inhaltlich und sprachlich manuell korrigierten Version verglichen wurde. Durch die KI ausgelassene Informationen wurden somit ergänzt, semantisch fehlerhafte Formulierungen wurden verbessert und unpassende Begriffe korrigiert. Diese Referenztexte wurden anschließend mit den KI-generierten Texten unter Verwendung des Bidirectional Encoder Representations from Transformers (BERT) - Scores verglichen. Dieser Score ist wiederum ein anerkanntes Maß für die Textähnlichkeit von maschinell erzeugten Übersetzungen, Zusammenfassungen und Texten.

Das BERT-Model generiert hochdimensionale Vektordarstellungen, bekannt als Embeddings. Diese Embeddings werden eingesetzt, um die Ähnlichkeit zwischen den Token, also den einzelnen Wörtern und Satzteilen, in verschiedenen Texten zu messen [78]. So werden beispielsweise die Sätze „Die Katze sitzt auf der Matte“ und „Ein Kätzchen ruht auf dem Teppich“ als semantisch ähnlich bewertet, da Begriffe wie „Katze“ und „Kätzchen“ sowie „Matte“ und „Teppich“ ähnliche Wortkonzepte darstellen. Durch die Analyse des gesamten Satzkontextes werden nicht nur Wortübereinstimmungen, sondern auch der Wortkontext berücksichtigt. Der BERT-Score quantifiziert diese Ähnlichkeit, indem er prüft, wie stark die kontextabhängigen Vektoren der beiden Sätze übereinstimmen – in diesem Fall würde ein hoher Score auf eine nahezu identische Bedeutung hinweisen, selbst wenn die Wortwahl variiert.

Der BERT-Score ermöglicht somit einen quantitativen Vergleich der semantischen Ähnlichkeit zwischen zwei Texten. Er setzt sich aus drei Hauptkomponenten zusammen: Precision (P), Recall (R) und F1-Score (F1).

-
- Die Precision bewertet den Anteil der Wörter im KI-generierten Text, die präzise zur Gesamtbedeutung im Vergleich zum Referenztext beitragen. Es geht also um die Relevanz und Genauigkeit der verwendeten Begriffe.
 - Der Recall misst den Informationsgehalt des generierten Textes. Es wird also genau gemessen, wie gut der KI-generierte Text alle relevanten Informationen des Referenztextes abdeckt, ohne wichtige Details auszulassen.
 - Der F1-Score bietet ein harmonisches Mittel aus Precision und Recall und gibt eine Gesamtbewertung, die sowohl die Vollständigkeit als auch die Genauigkeit des Textes berücksichtigt.

Ein hoher BERT-Score weist also auf eine hohe semantische Ähnlichkeit hin. Sind Texte identisch, so weist der F1-Score, der den Mittelwert zwischen dem Recall und der Precision bildet, einen Wert von 1.00 auf. Zusammenfassend dient der BERT-Score in dieser Studie als Werkzeug, die Qualität der KI-generierten Texte im Vergleich zu manuell erstellten Referenztexten zu bewerten. Die zusammengefassten Ähnlichkeits-Scores (Precision, Recall und F1), die auf einer Skala von 0 bis 1 angegeben werden, zeigen die allgemeine Textähnlichkeit an. Ein höherer BERT-Score weist auf eine größere Ähnlichkeit zwischen dem Referenztext und dem KI generierten Text und somit auf eine höhere Qualität der KI-generierten Ergebnisse hin [79].

2.4 Vereinfachung der KI-generierten Röntgenberichte

Um nun die Qualität und Verständlichkeit der Texte für die Patienten zu variieren, wurden die Texte mit zwei zusätzlichen Prompts durch die KI erneut generiert. Die Prompts, die hier genutzt worden sind, sollten verdeutlichen, wie einfache Befehlsvariationen im Sinne des Prompt-Engineering den Text signifikant abändern können. Der zweite Prompt wurde dementsprechend so formuliert, dass der initiale Text einen für den Patienten verständlicheren Text produziert:

- *„Schreibe den Röntgenbericht so um, dass er für einen Patienten verständlicher ist. Lasse dabei keine Informationen oder Inhalte aus.“*

Der dritte Prompt sollte mit einer einfachen Variation der gewählten Begriffe (Token) den Patiententext weiter vereinfachen.

- *„Schreibe den Röntgenbericht so um, dass er für Patienten jedes Bildungsstandes verständlich ist. Lasse dabei keine Informationen oder Inhalte aus.“*

Die auf diese Weise generierten Texte wurden dann erneut, wie bereits in Kapitel 2.3.2 beschrieben, mit dem FRE-Score und LIX hinsichtlich der Lesbarkeit analysiert. Darüber

hinaus wurden drei dem Durchschnitt entsprechende Texte der KI randomisiert ausgewählt und zur weiteren Evaluation durch Patienten bewertet.

2.5 Textevaluation durch Patienten

2.5.1 Patientenkollektiv

Die Patientenbefragung wurde in der Klinik für Mund-, Kiefer- und Gesichtschirurgie der Universitätsmedizin der Johannes Gutenberg-Universität Mainz durchgeführt. Die Patienten wurden hierfür zufällig ausgewählt und hinsichtlich einer Studienteilnahme aufgeklärt. Den teilnehmenden Patienten wurde zufällig einer der drei ausgewählten Texte gemeinsam mit dem zugehörigen Fragebogen ausgehändigt. Ausschlusskriterien zur Studienteilnahme stellten lediglich eine medizinische Vorbildung oder ein Alter unter 18 Jahren dar. Das Studienkollektiv sollte auf diese Weise eine möglichst allgemeine Patientengruppe widerspiegeln. Darüber hinaus wurden keine weiteren Ausschlusskriterien festgelegt, um eine breite und repräsentative Stichprobe des typischen Patientenkollektivs im klinischen Kontext einzubeziehen. Sowohl die Patienten als auch die Beobachter waren bezüglich der Gruppenzuordnung verblindet. Die verschiedenen Textversionen wurden anonymisiert, gemischt und randomisierten ausgehändigt, sodass weder die Teilnehmer noch die Beobachter wussten, welche Version jeweils bewertet wurde. Sämtliche Daten wurden anonym zwischen dem 1. Juni und dem 30. August 2024 erhoben, ohne dass persönliche Informationen erfasst wurden. Ein Rücktritt von der Studie oder der Widerruf des Einverständnisses waren jederzeit ohne Konsequenzen möglich. Insgesamt nahmen 150 Patienten an der Studie teil, wobei jeweils 50 Patienten eine der drei Textversionen bewerteten.

2.5.2 Fragebögen

Der verwendete Fragebogen enthielt 11 Fragen zur Evaluation von Verständlichkeit, Struktur und Informationsgehalt, Empathie und Ton sowie Motivation und zukünftiger Perspektive. Die Bewertung erfolgte auf einer fünfstufigen Likert-Skala von 1 („trifft voll zu“) bis 5 („trifft gar nicht zu“) mit den dazwischen liegenden Möglichkeiten „trifft eher zu“ (2), „weder noch“ (3) und „trifft eher nicht zu“ (4). Die jeweiligen Fragen sind in Abbildung 4 dargestellt.

Aussagen	Trifft voll zu	Trifft eher zu	Weder noch	Trifft eher nicht zu	Trifft gar nicht zu
Ich fand den Röntgenbericht leicht zu verstehen.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Die verwendeten Fachbegriffe im Röntgenbericht wurden klar erklärt.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Ich konnte den Befund des Röntgenberichts ohne zusätzliche Hilfe verstehen.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Die Struktur des Röntgenberichts hat mir geholfen, die Informationen leicht zu erfassen.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Die Informationen im Röntgenbericht waren ausreichend detailliert, um meine Fragen zu beantworten.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Der Ton des Röntgenberichts war respektvoll und berücksichtigte meine Situation als Patient.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Der Röntgenbericht vermittelte eine empathische Haltung gegenüber meinem Verständnisniveau.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Der Röntgenbericht hilft mir dabei, richtige Fragen an meinen Behandler stellen zu können.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Nach dem Lesen des Röntgenberichts kann ich mit meinem Behandler über meine Diagnose diskutieren und ein Gespräch auf Augenhöhe führen.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Der Röntgenbericht motiviert mich, mich in Zukunft mehr mit dem Thema Mundhygiene und Gesundheit auseinanderzusetzen.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Ich würde mir wünschen, dass alle meine medizinischen Berichte so verständlich wie dieser Röntgenbericht sind.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Abbildung 4: Fragebogen mit 11 Fragen zur Evaluation von Verständlichkeit, Struktur und Informationsgehalt, Empathie und Ton sowie Motivation und zukünftiger Perspektive der originalen KI-generierten und von ChatGPT vereinfachten Röntgenberichte. Die Bewertung erfolgte auf einer fünfstufigen Likert-Skala von 1 („trifft voll zu“) bis 5 („trifft gar nicht zu“) mit den dazwischen liegenden Möglichkeiten „trifft eher zu“ (2), „weder noch“ (3) und „trifft eher nicht zu“ (4).

2.6 Power Analyse und Fallzahlberechnung

Zur Fallzahlplanung wurde vor Studienbeginn eine a priori Power-Analyse durchgeführt. Diese gliederte sich entsprechend der Studienplanung in zwei Teile.

Zur Bewertung potenzieller Unterschiede in der Lesbarkeit zwischen KI-generierten Berichten und von Studierenden verfassten Texten wurden zwei Studien herangezogen, die die Lesbarkeit medizinischer Informationen untersucht haben. Hier zeigte sich ein signifikanter Unterschied von 11,6 Punkten in den FRE-Werten, wobei ChatGPT-Antworten niedrigere Durchschnittspunktzahlen (Mittelwert $34,9 \pm 11,2$) aufwiesen als medizinische Informationen auf Google-Webseiten (Mittelwert $46,5 \pm 14,3$) [80]. Ein ähnlicher Trend wurde auch für den LIX festgestellt, wobei ein Unterschied von 10 Punkten zwischen von Menschen verfassten und von ChatGPT generierten wissenschaftlichen Einleitungen nachgewiesen wurde [81]. Basierend auf diesen Ergebnissen und um eine Teststärke (Power) von 80 % bei Einhaltung eines Signifikanzniveaus von 5 % zu erreichen, sind mindestens 25 Stichproben pro Gruppe (Studienarm) erforderlich.

Hinsichtlich des zweiten Studienabschnitts, der Patientenbefragung, wurde die Stichprobengröße basierend auf einer kürzlich durchgeführten Studie durchgeführt. Die Arbeitsgruppe um Rajpurohit et al. zeigte im Rahmen der Evaluation eines Patienteninformationsflyers zur Leberzirrhose eine signifikante Verbesserung der Wissenswerte im Vorher-Nachher-Test [82]. Konkret wurde eine Verbesserung von 4 Punkten berichtet (Vorher-Test: Median = 12, Bereich: 6–20; Nachher-Test: Median = 16, Bereich: 12–22; $p < 0,05$), was einem moderaten Effekt (Cohen's $d = 0,4$) entspricht. Unter der Annahme einer ähnlichen Effektstärke ist zum Erreichen einer Teststärke (Power) von 80 % unter Beibehaltung eines Signifikanzniveaus von 5 % eine Stichprobengröße von mindestens 50 Probanden pro Gruppe (insgesamt 150 Teilnehmer) notwendig.

2.7 Statistische Analyse

Die zur statistischen Analyse verwendeten Softwarepakete waren GraphPad Prism 9.0 (GRAPHPAD SOFTWARE, LLC, Boston, USA), G*Power 3.1 (Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Deutschland) und Excel 16.76 (Microsoft Corporation, Redmond, USA). Alle Daten werden als Mittelwert +/- Standardabweichung (SD) präsentiert. Das Signifikanzniveau wurde bei $p < 0,05$ festgesetzt. Unterschiede zwischen den von Studierenden verfassten und KI-generierten Texten wurden mit einem zweiseitigen Student's t-Test analysiert. Eine nachfolgende post hoc Power-Analyse wurde für jeden Test durchgeführt, um die durch den t-Test erreichte Power zu überprüfen. Für die Auswertung der durch die Patienten ausgefüllten Fragebögen mittels Likert-Skala zur Analyse des Textverständnisses der KI-generierten Röntgenberichte wurden dieselben oben

beschriebenen Softwarepakete verwendet. Um potenzielle Unterschiede in der Verständlichkeit zwischen den KI-generierten Berichten zu bewerten, wurde eine ANOVA-Analyse durchgeführt, gefolgt von Tukey's Multiple Comparisons Test. Die Signifikanz wurde hier ebenfalls bei $p < 0,05$ festgesetzt und alle Daten sind als Mittelwert \pm Standardabweichung dargestellt.

2.8 Ethikvotum

Diese Studie wurde unter Einhaltung aller ethischen Richtlinien durchgeführt, wobei vor Beginn der Teilnahme von allen Teilnehmern mündlich eine informierte Einwilligung eingeholt wurde. Alle Daten wurden anonymisiert erhoben und ausgewertet. Nach Begutachtung durch die Ethikkommission der Landesärztekammer Rheinland-Pfalz, Mainz, Deutschland war keine weitere Genehmigung erforderlich, da die Studie in erster Linie als Maßnahme zur Qualitätssicherung konzipiert war und die Befragung anonym durchgeführt wurde (Antragsnummer: 2024-17526).

3 Ergebnisse

Im folgenden Abschnitt werden die Ergebnisse der Analyse der KI-generierten sowie von Studierenden verfassten Röntgenberichte vorgestellt. Zu Beginn erfolgte hierfür die Bewertung von Textqualität, Lesbarkeit und Verständlichkeit anhand verschiedener Sprachparameter sowie die Evaluation der inhaltlichen Vollständigkeit und grammatikalischen Korrektheit der Texte. Anschließend erfolgte die Auswertung der Patientenbewertungen im Hinblick auf die Verbesserung des Verständnisses und der Patientenkommunikation. Während der gesamten Studie nutzten die Studierenden die maximale Bearbeitungszeit sowohl zum manuellen Verfassen der Röntgenberichte (30 Minuten) als auch zum Ausfüllen der Checkboxliste (10 Minuten) konsequent aus. Trotz ihrer formalen Ähnlichkeit zu den Referenztexten zeigten die KI-generierten Berichte einen signifikanten Informationsverlust. Demgegenüber erwiesen sich die von ChatGPT vereinfachten Versionen als effektiv in der Steigerung der Lesbarkeit und Verständlichkeit, was insbesondere in der Bewertung durch die Patienten zum Ausdruck kam.

3.1 Röntgenberichte - ChatGPT vs. Studierende

3.1.1 Zeitaufwand zur Erstellung der Röntgenberichte

Im Rahmen der Untersuchung zeigte sich, dass alle teilnehmenden Studierenden für die freie Ausformulierung eines Röntgenberichts im Fließtext die ihnen zur Verfügung stehende Zeit von 30 Minuten vollständig ausschöpften. Im Vergleich dazu betrug der zeitliche Aufwand für das Ausfüllen der standardisierten Checkboxliste zehn Minuten.

3.1.2 Lesbarkeit der Röntgenberichte

Die statistische Analyse ergab keine signifikanten Unterschiede in der Lesbarkeit zwischen den KI-generierten Texten und den durch Studierende verfassten Berichten. Die durchschnittliche Lesbarkeit war für beide Gruppen ähnlich (ChatGPT: $50,55 \pm 7,80$; Studierende: $51,19 \pm 5,02$; $p = 0,4911$; $t = 0,6898$; $df = 198$; Abbildung 5). Bei der individuellen Untersuchung jedes Semesters, dargestellt in Abbildung 5 B, zeigte der FRE-Score signifikante Unterschiede, wobei KI-generierte Texte im Vergleich zu Texten von Studierenden aus dem ersten klinischen Semester eine schlechtere Lesbarkeit aufwiesen ($56,65 \pm 6,70$ vs $49,6 \pm 7,17$; $p = 0,002$; $t = 3,213$; $df = 38$). Im Vergleich zum dritten ($47,14 \pm 6,97$ vs $51,14 \pm 4,55$; $p = 0,033$; $t = 2,203$; $df = 40$) und vierten ($46,35 \pm 7,29$ vs $52,15 \pm 5,08$; $p = 0,006$; $t = 2,918$; $df = 38$) klinischen Semester war die Lesbarkeit der KI-generierten Texte jedoch signifikant besser, wobei kein signifikanter Unterschied zum zweiten ($p = 0,3929$; $t = 0,8647$; $df = 36$) und fünften ($p = 0,123$; $t = 1,577$; $df = 38$) klinischen Semester bestand (Abbildung 5 B).

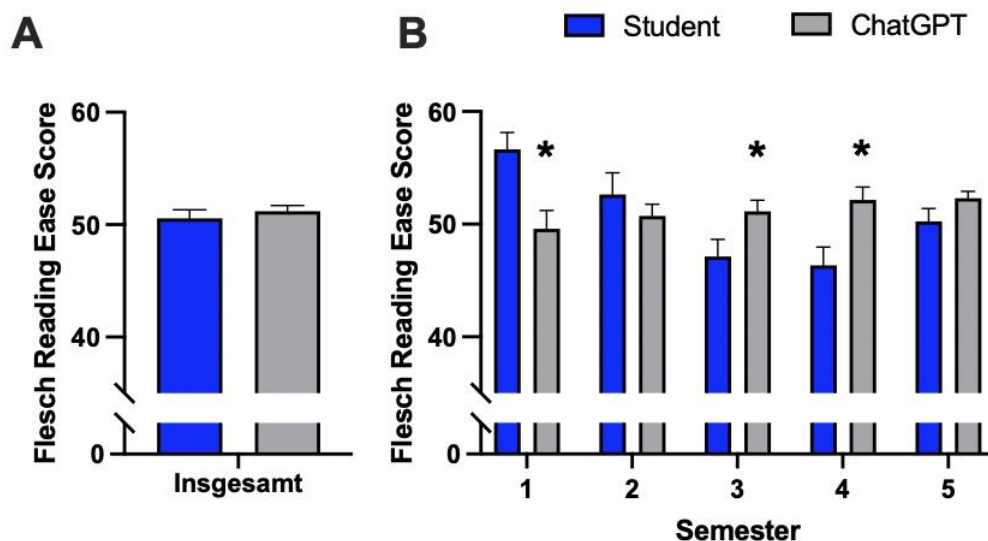


Abbildung 5: Lesbarkeit von KI-generierten Röntgenberichten im Vergleich zu studentisch verfassten radiologischen Berichten semesterunabhängig (A) und semesterabhängig (B), bewertet mit dem FRE-Score. Die Daten repräsentieren den Mittelwert +/- Standardabweichung. Stichprobengröße: $n = 100$ (A); Semester 1: $n=20$, Semester 2: $n=19$, Semester 3: $n=21$, Semester 4: $n=20$, Semester 5: $n=20$ (B); * = $p < 0,05$.

Neben dem FRE-Score wurde weiterhin der LIX für alle Texte ermittelt. Wie in Abbildung 6 A dargestellt, zeigte sich kein signifikanter Unterschied in der Lesbarkeit zwischen den beiden Gruppen ($48,98 \pm 5,0$ gegenüber $48,0 \pm 2,85$; $p = 0,091$; $t = 1,699$; $df = 198$). Der LIX zeigt über die verschiedenen Semester hinweg ähnliche Tendenzen wie der FRE-Score, jedoch mit gegensätzlichen Werten, wie in Abbildung 6 B dargestellt. Während ein hoher LIX-Wert auf langen Sätzen und komplexen Wörtern basiert, entspricht dies einem niedrigen FRE-Wert, welcher mit der Anzahl an kurzen Sätzen und einfachen Wörtern ansteigt. Somit verhalten sich die beiden Metriken gegensätzlich zueinander. Signifikante Unterschiede in der Lesbarkeit wurden für das erste ($46,27 \pm 4,0$ vs $48,81 \pm 3,44$; $p = 0,037$; $t = 2,157$; $df = 38$), das dritte ($51,64 \pm 4,89$ vs $48,01 \pm 2,84$; $p = 0,005$; $t = 2,944$; $df = 40$) und das vierte Semester ($50,67 \pm 4,68$ vs $47,32 \pm 2,90$; $p = 0,0098$; $t = 2,719$; $df = 38$) festgestellt. Für das zweite ($p = 0,39$; $t = 0,877$; $df = 36$) und fünfte Semester ($p = 0,151$; $t = 1,464$; $df = 38$) wurden hingegen keine signifikanten Unterschiede in der Lesbarkeit ermittelt (Abbildung 6 A und B).

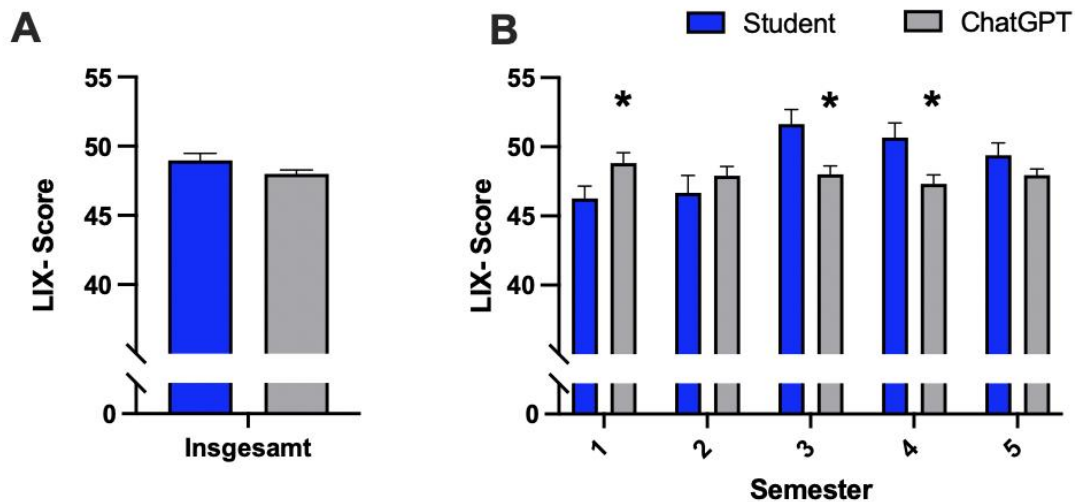


Abbildung 6: Lesbarkeit von KI-generierten Röntgenberichten im Vergleich zu von Studierenden verfassten radiologischen Berichten semesterunabhängig (A) und semesterabhängig (B), bewertet mit dem LIX-Score. Die Daten repräsentieren den Mittelwert +/- Standardabweichung. Stichprobengröße: $n = 100$ (A); Semester 1: $n=20$, Semester 2: $n=19$, Semester 3: $n=21$, Semester 4: $n=20$, Semester 5: $n=20$ (B); * = $p < 0,05$.

3.1.3 Ähnlichkeit und Informationsgenauigkeit der Röntgenberichte

Der BERT-Score wurde zum Vergleich der KI-generierten Röntgenberichte mit den Referenztexten erhoben. Wie in Abbildung 7 A dargestellt zeigt sich eine große Ähnlichkeit zwischen den Referenztexten und den Originaltexten, quantifiziert durch eine hohe Präzision (P) von $0,967 \pm 0,036$, einen Recall (R) von $0,958 \pm 0,037$ und einen F1-Score von $0,962 \pm 0,036$. Dennoch offenbarte die inhaltliche Analyse einen signifikanten Mangel an relevanten Informationen in den KI-generierten Texten, die nicht aus den Checkboxlisten übernommen wurden. Dieser Unterschied im Informationsgehalt zwischen den von den Studierenden diagnostizierten Befunden in den Checkboxlisten und den daraus generierten Röntgenberichten stellte sich als statistisch signifikant dar (Abbildung 7 B). Die Studierenden diagnostizierten insgesamt $44,6 \pm 6,0$ Befunde in der Checkboxliste, während die KI nur $41,3 \pm 7,0$ Befunde in die Röntgenberichte übernahm ($p = 0,0004$; $t = 3,586$; $df = 198$). Im Detail zeigte sich, wie in Abbildung 7 C dargestellt, kein signifikanter Unterschied für das erste Tabellenblatt ($8,53 \pm 1,06$ vs $8,55 \pm 1,02$; $p = 0,892$; $t = 0,1361$; $df = 198$), jedoch schloss ChatGPT signifikant weniger Befunde aus dem zweiten ($23,03 \pm 3,67$ vs $21,22 \pm 4,92$; $p = 0,0035$; $t = 2,951$; $df = 198$) und dritten Tabellenblatt ($13,07 \pm 4,40$ vs $11,56 \pm 4,23$; $p = 0,0141$; $t = 2,476$; $df = 198$) in den Röntgenbericht ein.

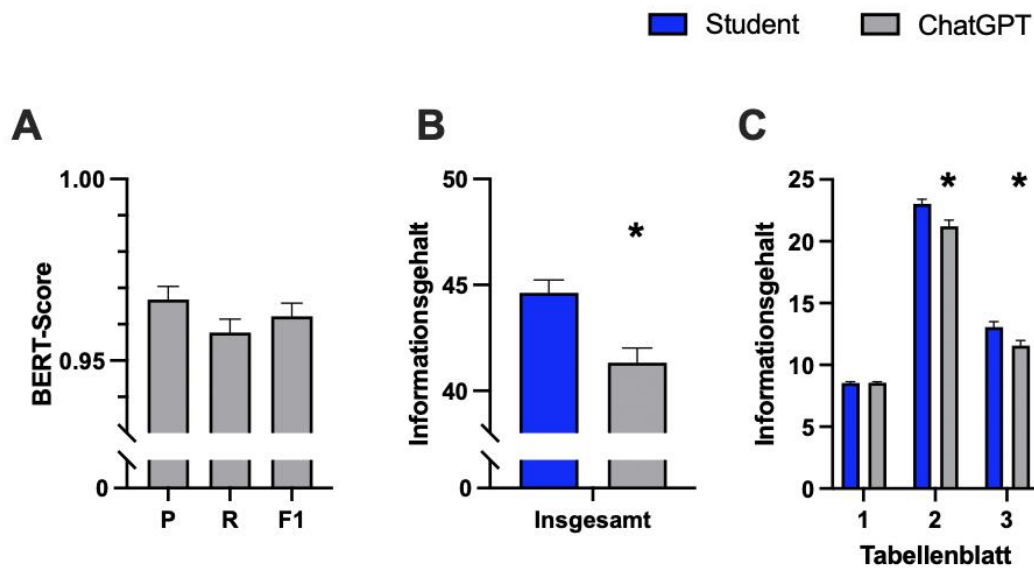


Abbildung 7: Bewertung der Ähnlichkeit der KI-generierten Röntgenberichte im Vergleich zu Referenztexten unter Verwendung des BERT-Scores (A) mit Precision (P), Recall (R) und F1-Score (F1). Die Genauigkeit der KI-generierten radiologischen Berichte wurde weiterhin über die enthaltenen diagnostischen Informationen für die gesamte Checkboxliste (B) sowie individuell für jedes Arbeitsblatt der Checkboxliste bewertet. Die Daten repräsentieren den Mittelwert +/- Standardabweichung. Stichprobengröße: $n = 100$. * = $p < 0,05$.

3.1.4 Quantitative Analyse der sprachlichen Merkmale und der Fehlerquote

Darüber hinaus wurden die Wortanzahl, Satzanzahl, Silbenzahl, Diphthong und Zeichenanzahl der KI-generierten Texte im Vergleich zu denen der von den Studierenden formulierten Röntgenberichte erhoben. Dabei zeigten die von ChatGPT-generierten radiologischen Befundtexte eine signifikante Wortzahlreduktion um etwa 24% ($265,6 \pm 95,4$ vs. $200,6 \pm 37,3$ Wörter; $p < 0,0001$; $t = 6,347$; $df = 198$) und der Satzanzahl ($p = 0,007$; $t = 2,726$; $df = 198$), begleitet von signifikanten Verringerungen der Silben ($p < 0,0001$; $t = 6,823$; $df = 198$), Diphthonge ($p < 0,0001$; $t = 8,643$; $df = 198$) und Zeichen ($p < 0,0001$; $t = 6,841$; $df = 198$; Abbildung 8 A – E).

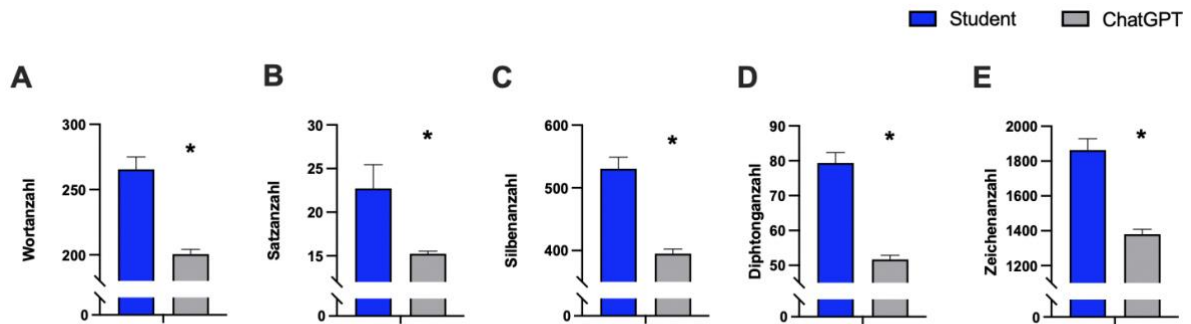


Abbildung 8: Analyse der Wort- (A), Satz- (B), Silben- (C), Diphthong- (D) und Zeichenanzahl (E) von KI-generierten im Vergleich zu von Studierenden verfassten Röntgenberichten. Die Daten repräsentieren den Mittelwert +/- Standardabweichung. Stichprobengröße: $n = 100$. * = $p < 0,05$.

Wie in Abbildung 9 dargestellt, zeigten die von ChatGPT generierten radiologischen Berichte eine signifikante Reduktion der durchschnittlichen Satzlänge im Vergleich zu den von Studierenden verfassten Berichten (A: $14,5 \pm 3,1$ vs. $13,7 \pm 1,8$ Wörter; $p = 0,0462$; $t = 2,007$; $df = 198$). Hinsichtlich des Anteils an langen Wörtern wurde jedoch kein Unterschied zwischen den Gruppen festgestellt (B: $p = 0,6112$; $t = 0,509$; $df = 198$).

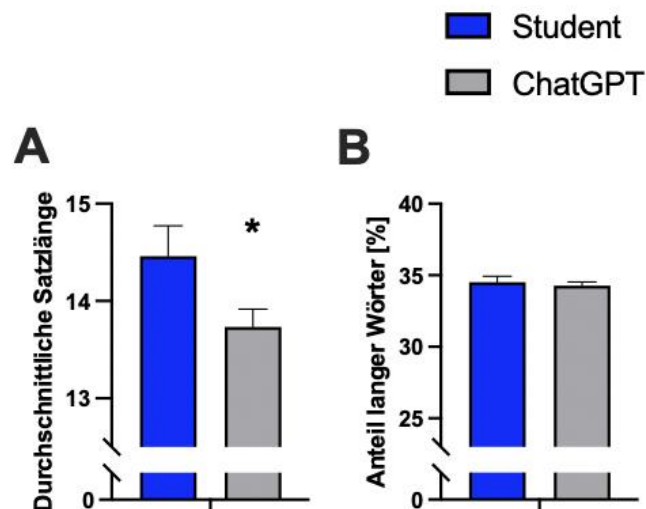


Abbildung 9: Analyse der Satzlänge und des Anteils langer Wörter (mehr als sechs Buchstaben) in KI-generierten Röntgenberichten im Vergleich zu von Studierenden verfassten Berichten. Die Daten repräsentieren den Mittelwert +/- Standardabweichung. Stichprobengröße: $n = 100$. * $p < 0,05$ vs. Studierende.

Abschließend wurden die von den Studierenden formulierten Befunde im Hinblick auf Rechtschreib-, Grammatik- und Interpunktionsfehler analysiert, sowie der Fehlerquotient berechnet. Die Abbildung 10 A und B zeigen, dass die KI-generierten Texte im Gegensatz zu den von Studierenden verfassten Berichte keine Rechtschreib-, Grammatik- oder

Zeichensetzungsfehler aufwiesen (A: $7,7 \pm 7,2$ vs. 0; $p < 0,0001$; $t = 10,59$; $df = 198$; B: $2,9 \pm 2,9$ vs. 0; $p < 0,0001$; $t = 10,41$; $df = 198$).

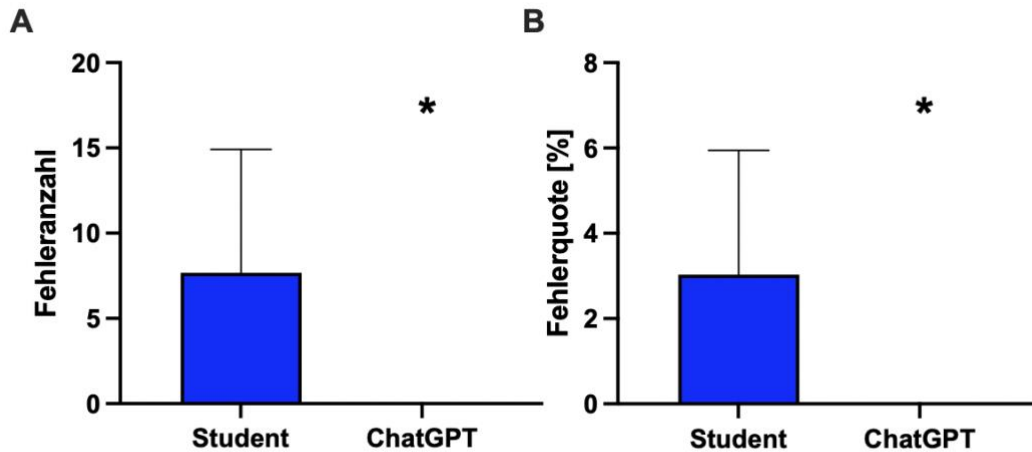


Abbildung 10: Analyse der Fehleranzahl (A), einschließlich Grammatik, Rechtschreibung und Zeichensetzung, sowie des Fehlerquotienten (B), berechnet als Fehler / Wörter * 100, von KI-generierten Röntgenberichten im Vergleich zu von Studierenden verfassten Berichten. Die Daten repräsentieren den Mittelwert \pm Standardabweichung. Stichprobengröße: $n = 100$. * = $p < 0,05$.

3.2 Auswertung der von ChatGPT vereinfachten Röntgenberichte

Die ursprünglich von ChatGPT erstellten Texte wurden mithilfe von zwei zusätzlichen Prompts überarbeitet und sprachlich vereinfacht. Das Ziel dieser Überarbeitung war die Verbesserung der Lesbarkeit, um die Texte für Patienten verständlicher zu machen. Diese neu generierten Texte wurden wie bereits im ersten Studienabschnitt hinsichtlich ihrer Lesbarkeit sowie Sprache bewertet. Im Folgenden bezeichnet Prompt 1 die originale KI-generierte Version des Röntgenberichts, während Prompt 2 für die erste und Prompt 3 für die zweite Vereinfachung der Texte verwendet wurden.

3.2.1 Lesbarkeit der vereinfachten Röntgenberichte

Zur Bewertung der Lesbarkeit dieser vereinfachten Texte kam erneut der FRE-Score zur Anwendung. Abbildung 11 zeigt eine Gegenüberstellung des FRE-Scores der durch die Studierenden formulierten Befundberichte und jener Texte, die mithilfe der drei verschiedenen Prompts generiert worden sind. Dabei werden sowohl die originalen KI-generierten Texte als auch die beiden von ChatGPT vereinfachten Versionen in den Vergleich miteinbezogen. Die Ergebnisse zeigen einen signifikanten Einfluss des Semesters und somit des Weiterbildungsstandes auf die Lesbarkeit (A: $F(4,96) = 2,521$; $p = 0,0461$). Weiterhin offenbarte die statistische Analyse eine signifikante Zunahme der Lesbarkeit je stärker die

Röntgenberichte vereinfacht wurden ($A: F(2,213) = 32,80; p < 0,0001$). Mit Ausnahme des ersten und zweiten Semesters waren die von Studierenden verfassten Berichte gleichzeitig die am schwierigsten zu verstehenden, während die Erstellung der Röntgenberichte durch KI, die erste Vereinfachung und schließlich die zweite Vereinfachung in graduell ansteigenden FRE-Werten resultierten. Bei isolierter Betrachtung der von Studierenden formulierten Texte ließ sich ein rückläufiger Trend in der Lesbarkeit nachweisen. Der FRE-Score nahm von $56,6 \pm 6,7$ im ersten klinischen Semester auf $50,3 \pm 5,1$ im fünften klinischen Semester signifikant ab. Die mit dem originalen Prompt generierten Texte erreichten einen FRE-Wert von $49,6 \pm 7,2$ im ersten und $52,3 \pm 2,7$ im fünften klinischen Semester. Die erste Vereinfachung mit dem zweiten Prompt zeigte eine Verbesserung des FRE-Scores von $53,9 \pm 3,7$ im ersten klinischen Semester auf $54,9 \pm 5,5$ im fünften klinischen Semester. Die höchste Lesbarkeit wurde durch die stärkste Vereinfachung mit dem dritten Prompt erzielt, dessen Texte einen FRE-Score von $55,6 \pm 5,1$ im ersten und $56,2 \pm 4,0$ im fünften klinischen Semester aufwiesen.

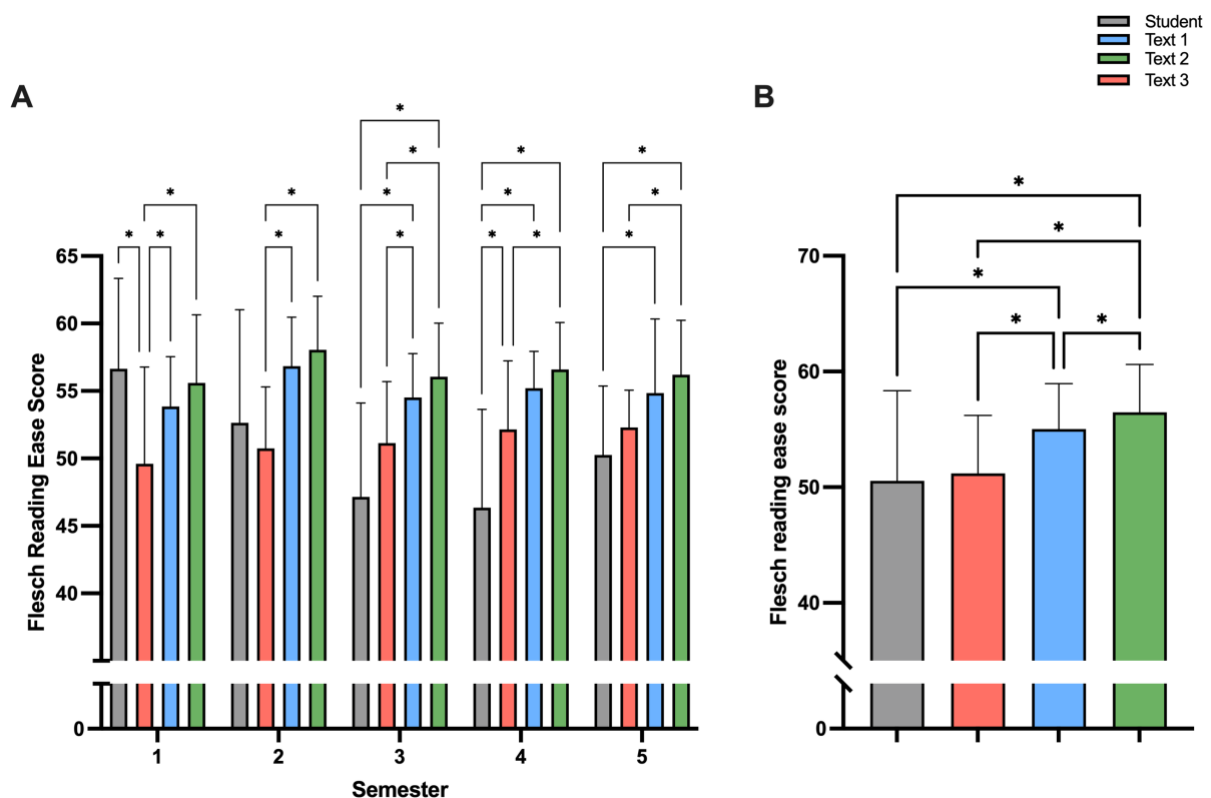


Abbildung 11: Analyse der FRE-Werte der von KI-generierten und von KI-vereinfachten Röntgenberichte im Vergleich zu von Studierenden verfassten Berichten insgesamt (A) und einzeln für jedes Semester (B), bewertet mit dem FRE-Score. Dabei wurden die KI-generierte Berichte mit zwei weiteren Prompts (Prompt 2 und Prompt 3) in vereinfachte Versionen umformuliert, um die Lesbarkeit zu verbessern. Die Daten repräsentieren den Mittelwert \pm Standardabweichung. Stichprobengröße: $n = 100$; Semester 1: $n=20$, Semester 2: $n=19$, Semester 3: $n=21$, Semester 4: $n=20$, Semester 5: $n=20$. * = $p < 0,05$.

Fasst man, wie in Abbildung 11 B dargestellt, alle Semester zusammen, so festigt sich der oben beschriebene Trend der zunehmenden Lesbarkeit bei schrittweiser Vereinfachung der Texte. Die von Studierenden formulierten Röntgenberichte wiesen im Vergleich zu den drei KI-generierten sowie KI-vereinfachten Versionen statistisch signifikant geringere FRE-Werte und somit eine schlechtere Lesbarkeit auf ($F(2, 207) = 29,41$; $df = 3$; $p < 0.0001$). Die mithilfe des ersten Prompts generierten Texte erreichten einen FRE-Wert von $50,1 \pm 5,8$, was im Vergleich zu den studentischen Texten auf eine etwas bessere Lesbarkeit hindeutet. Prompt zwei zeigte eine Lesbarkeit von $52,6 \pm 4,9$. Die mithilfe des dritten Prompts am stärksten vereinfachte Textversion zeigte gleichzeitig den höchsten FRE-Wert ($54,3 \pm 3,6$).

3.2.2 Quantitative Analyse der sprachlichen Merkmale der vereinfachten Röntgenberichte

Wie im ersten Studienabschnitt wurden weiterhin die Satzanzahl, die Wortanzahl und der Anteil langer Wörter der von ChatGPT generierten Röntgenberichte evaluiert. Die studentischen Texte zeigen hierbei eine höhere durchschnittliche Satzanzahl (Abbildung 12A: $21,4 \pm 5,6$) im Vergleich zu den verschiedenen KI-generierten Textversionen. Die durch Prompt 1 erstellten Texte enthielten $15,2 \pm 4,3$ Sätze, während Prompt 2 mit $16,5 \pm 3,8$ Sätzen und Prompt 3 mit $17,1 \pm 4,1$ Sätzen eine leichte Zunahme der Satzanzahl aufwiesen. Die durchschnittliche Wortanzahl der studentischen Texte lag bei $286,7 \pm 45,2$ Wörtern. Die durch Prompt 1 generierten Texte zeigten hingegen eine deutlich geringere Wortanzahl von $214,5 \pm 37,6$ Wörtern, während Prompt 2 und Prompt 3 mit $236,8 \pm 40,1$ und $244,2 \pm 38,9$ Wörtern eine leichte Steigerung aufwiesen (Abbildung 12 A). Der Anteil langer Wörter zeigte in den studentischen Texten mit $36,4 \pm 7,5$ % einen leicht höheren Wert als die KI-generierten Texte. Prompt 1 enthielt $31,2 \pm 6,3$ %, Prompt 2 $32,7 \pm 5,8$ % und Prompt 3 $33,5 \pm 5,1$ % lange Wörter, was auf eine leichte Zunahme langer Wörter über die Prompt-Versionen hinweg hindeutet (Abbildung 12 C).

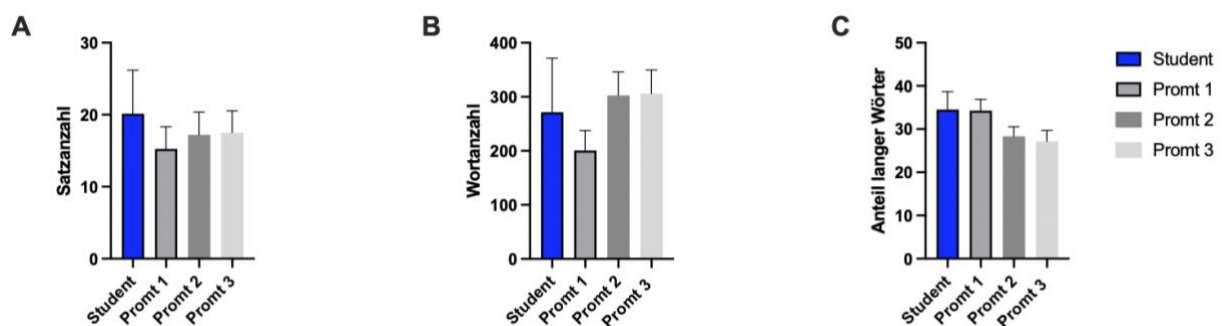


Abbildung 12: Analyse der durchschnittlichen Satzanzahl (A), Wortanzahl (B) und des Anteils langer Wörter (C) der von KI-generierten und von KI-vereinfachten Röntgenberichte im Vergleich zu von Studierenden verfassten Berichten. Dabei wurden die KI-generierte Berichte mit zwei weiteren Prompts (Prompt 2 und Prompt 3) in vereinfachte Versionen umformuliert, um die Lesbarkeit zu verbessern. Die Daten repräsentieren den Mittelwert \pm Standardabweichung. Stichprobengröße: $n = 100$. * = $p < 0,05$.

3.3 Patientenevaluation der KI-generierten Röntgenberichte

Im Rahmen der vorliegenden Studie wurden die mithilfe von KI generierten und vereinfachten Röntgenberichte von insgesamt 150 Patienten bewertet. Dabei erhielt jede der drei Gruppen (n=50) randomisiert eine Version des Röntgenberichts und bewertete die Texte auf Klarheit und Deutlichkeit, Struktur und Informationsgehalt, Ton und Empathie, Anleitung und Motivation sowie zukünftige Perspektive der Röntgenberichte anhand einer fünfstufigen Likert-Scala anhand 11 ausgewählter Fragen. Den ersten Text stellte die initiale KI-generierte Röntgenberichtversion dar, die durch den ursprünglichen Prompt (Prompt 1 / Text 1) generiert worden ist. Beim zweiten Röntgenbericht handelte es sich um die erste Vereinfachung, die den Text verständlicher für Patienten formulieren sollte (Prompt 2 / Text 2), während der dritte Röntgenbericht (Prompt 3 / Text 3) für jeden Patient unabhängig seines Bildungsniveaus gut verständlich und lesbar sein sollte. Die Auswertung der Ergebnisse der Bewertung durch die Patienten offenbarte signifikante Unterschiede zwischen den Textversionen in allen untersuchten Parametern.

3.3.1 Klarheit und Deutlichkeit der Röntgenberichte

Die Frage nach der Klarheit (Erklärungen der Fachbegriffe im Röntgenbericht) zeigte signifikante Unterschiede zwischen den Gruppen ($F(2,147) = 48,91$; $df = 2$; $p < 0.0001$; Abbildung 13 A). Dabei wurde Text 1 signifikant schlechter bewertet als Text 2 und 3. Die beiden vereinfachten Versionen (Text 2 und 3) unterschieden sich jedoch nicht signifikant voneinander. Ein ähnliches Muster zeigte sich bei der Frage, ob der Befund des Röntgenberichts ohne zusätzliche Hilfe verstanden werden konnte (Abbildung 13 B). Auch hier zeigten sich signifikante Unterschiede zwischen den Gruppen ($F(2,147) = 43,85$; $df = 2$; $p < 0.0001$), wobei der post hoc Vergleich ergab, dass Text 1 signifikant schlechter bewertet wurde als Text 2 und Text 3. Kein signifikanter Unterschied bestand zwischen den vereinfachten Röntgenberichten (Text 2 und 3).

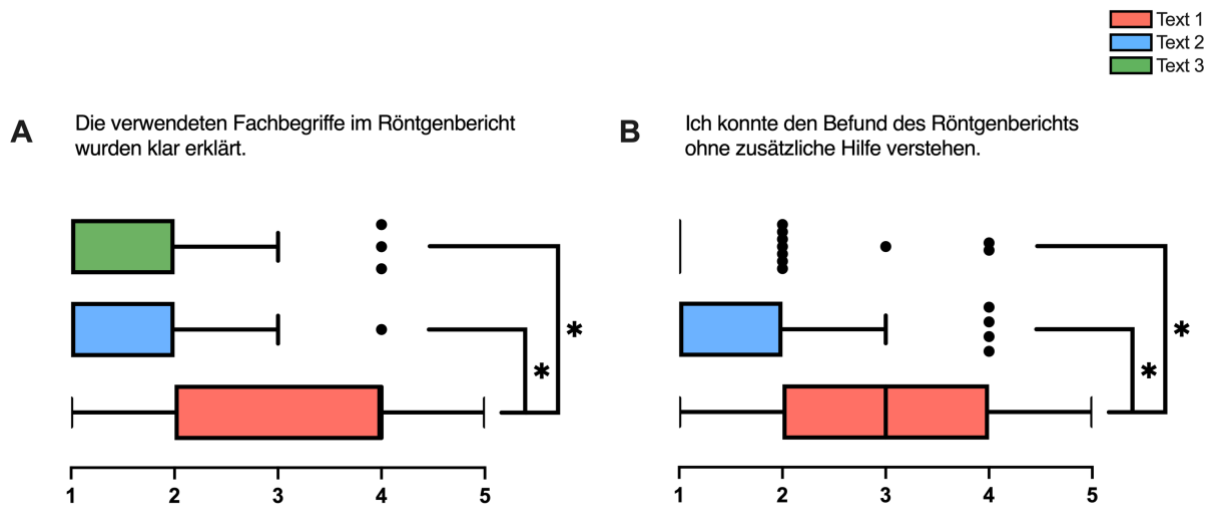


Abbildung 13: Bewertungen des Patientenverständnisses der Röntgenbefunde für die originalen KI-generierten Röntgenberichte (Text 1 – rot), die vereinfachte Version (Text 2 – blau) sowie die stark vereinfachte Version (Text 3 – grün). Evaluiert wurden die Erklärungen der verwendeten Fachbegriffe (A) sowie die Verständlichkeit ohne zusätzliche Hilfe (B). Stichprobengröße: $n=50$; * = $p < 0,05$.

3.3.2 Struktur und Informationsgehalt der Röntgenberichte

Bei der Frage, wie gut die Struktur des Röntgenberichts den Patienten hilft, die Informationen zu verstehen, zeigten sich signifikante Unterschiede ($F(2,147) = 30,91$; $df = 2$; $p < 0,0001$) zwischen den Gruppen (Abbildung 14 A). Der anschließende Vergleich mittels Tukey's Multiple Comparisons Test zeigte, dass Text 1 signifikant schlechter bewertet wurde als die Texte 2 und 3. Kein Unterschied bestand zwischen Text 2 und 3. Vergleichbare Ergebnisse lieferte die Auswertung der Frage, ob der Röntgenbericht ausreichend detailliert war, um alle Fragen zu beantworten ($F(2,147) = 31,53$; $df = 2$; $p < 0,0001$; Abbildung 15 B).

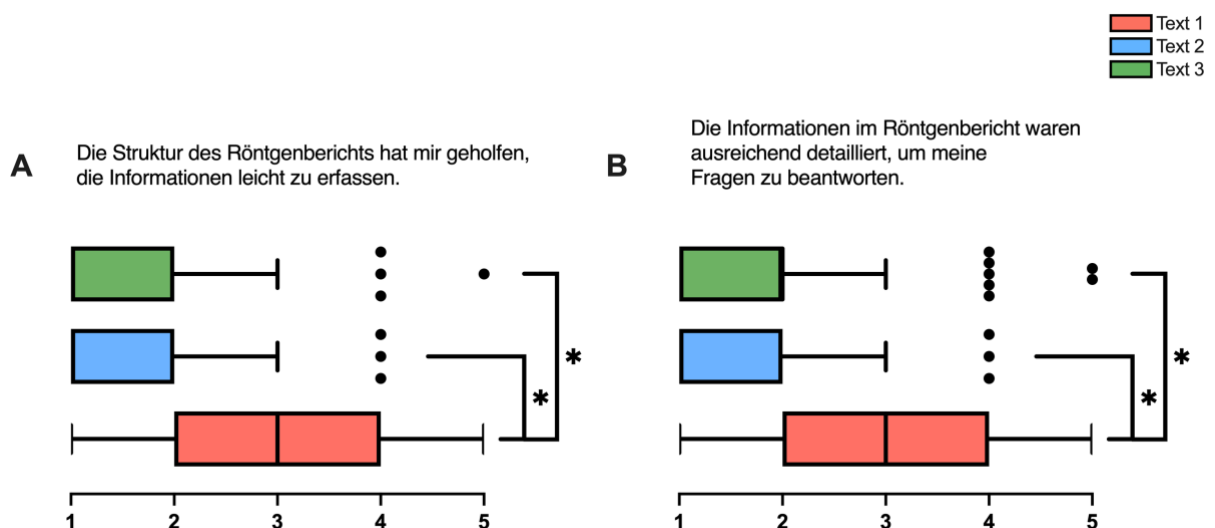


Abbildung 14: Bewertungen des Patientenverständnisses der Röntgenbefunde für die originalen KI-generierten Röntgenberichte (Text 1 – rot), die vereinfachte Version (Text 2 – blau) sowie die stark vereinfachte Version (Text 3 – grün). Evaluiert wurden die Struktur des Röntgenberichts (A) sowie die Detailliertheit der vermittelten Informationen (B). Stichprobengröße: $n=50$; * = $p < 0,05$.

3.3.3 Ton und Empathie des Röntgenberichts

Außerdem wurden die Patienten dazu befragt, ob der Ton des Röntgenberichts als respektvoll empfunden wurde (Abbildung 15 A) und ob ausreichend Empathie für ihr Verständnisniveau gezeigt wurde (Abbildung 15 B). Zwischen den drei Texten ließ sich hierbei ein signifikanter Unterschied feststellen (A: $F(2,147) = 11,42$; $df = 2$; $p < 0.0001$). Wie in den vorangegangenen Fragen offenbarte die post hoc Analyse eine bessere Bewertung zugunsten der Texte 2 und 3 im Vergleich zu Text 1, jedoch ohne signifikanten Unterschied zwischen Text 2 und 3. Vergleichbare Ergebnisse zeigten sich ebenso hinsichtlich der Verwendung einer empathischen Formulierung (B: $F(2,147) = 28,62$; $df = 2$; $p < 0.0001$).

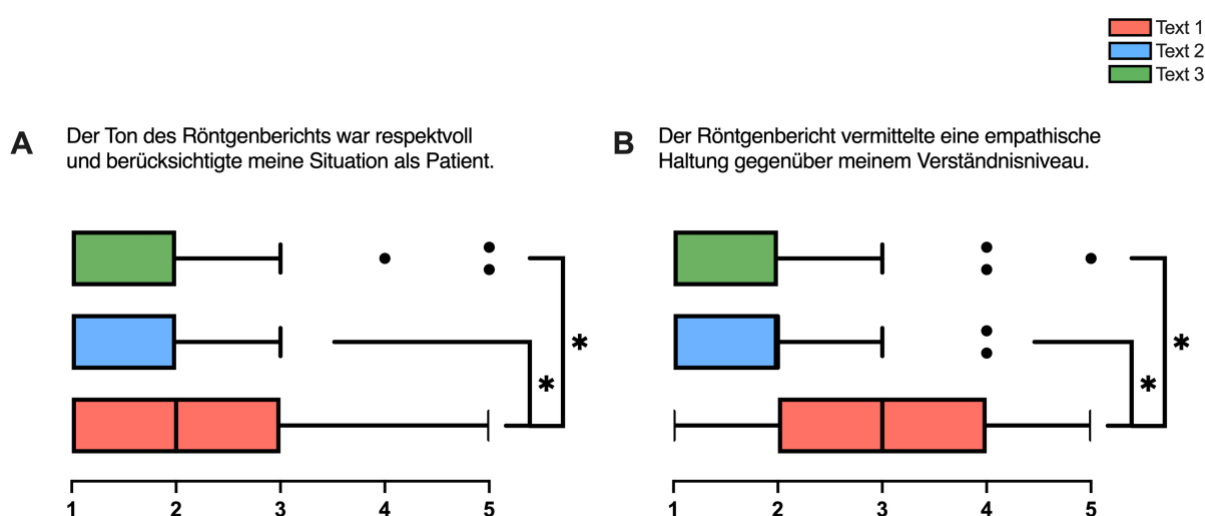


Abbildung 15: Bewertungen des Patientenverständnisses der Röntgenbefunde für die originalen KI-generierten Röntgenberichte (Text 1 – rot), die vereinfachte Version (Text 2 – blau) sowie die stark vereinfachte Version (Text 3 – grün). Ton des Röntgenberichts (A), sowie die empathische Haltung gegenüber dem Verständnisniveau (B). Stichprobengröße: $n=50$; * = $p < 0,05$.

3.3.4 Anleitung und Motivation der Röntgenberichte

Um das Patientenverständnis weiter zu beurteilen, beantworteten die Patienten zwei Fragen (Abbildung 16) in denen sie angeben sollten, wie gut sie sich nach dem Lesen des Röntgenberichts in der Lage fühlten, Fragen an den Behandler zu stellen und eine informierte Diskussion über die Diagnose zu führen. Für beide Fragen zeigten sich ähnliche Ergebnisse mit wiederum signifikanten Unterschieden versionsunabhängig (A: $F(2,147) = 12,23$; $df = 2$; $p < 0.0001$; B: $F(2,147) = 26,03$; $df = 2$; $p < 0.0001$). Erneut wurden die Texte 2 und 3 besser bewertet als Text 1, jedoch ohne Unterschied zwischen den vereinfachten Versionen.

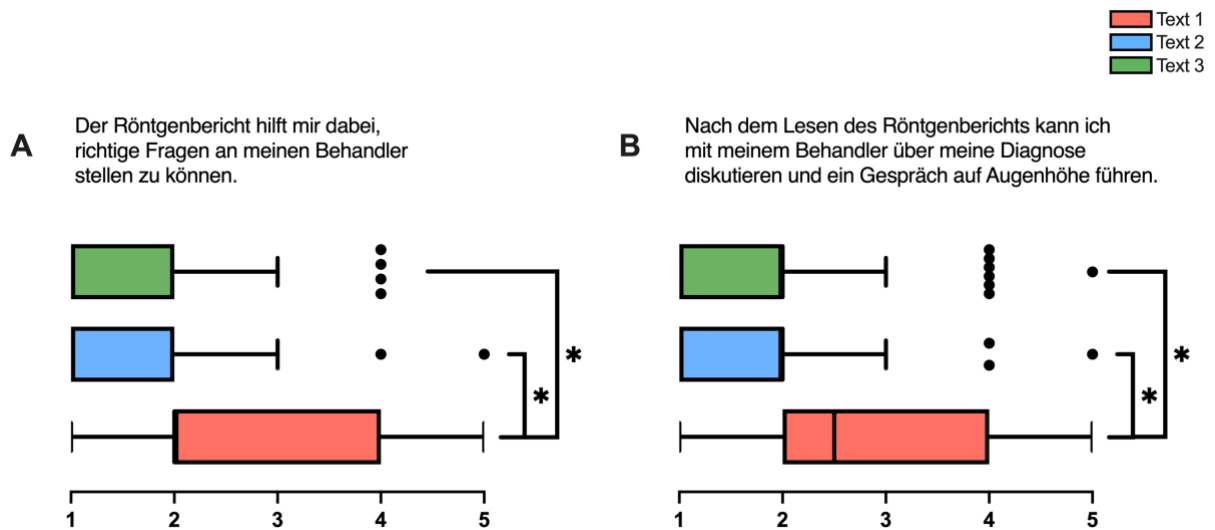


Abbildung 16: Bewertungen des Patientenverständnisses der Röntgenbefunde für die originalen KI-generierten Röntgenberichte (Text 1 – rot), die vereinfachte Version (Text 2 – blau), sowie die stark vereinfachte Version (Text 3 – grün). Evaluiert wurde, ob der Röntgenbericht hilft in weiteren Gesprächen die richtigen Fragen zu stellen (A), sowie ob die Möglichkeit ein weiteres Gespräch auf Augenhöhe zu führen besteht (B). Stichprobengröße: $n=50$; $* = p < 0,05$.

Wie in Abbildung 17 gezeigt, fanden die befragten Patienten beide vereinfachten Versionen des Röntgenberichts verständlicher als den ursprünglichen KI-generierten Bericht. Darüber hinaus motivierten die bevorzugten Textversionen die Befragten sich in Zukunft mit dem Thema Mundhygiene und Gesundheit auseinanderzusetzen (A: $F(2,147) = 51,61$; $df = 2$; $p < 0.0001$; B: $F(2,147) = 8,87$; $df = 2$; $p < 0.0001$). Für beide Fragen zeigten sich signifikant bessere Bewertungen für die Texte 2 und 3 im Vergleich zu Text 1, wobei sich die beiden vereinfachten Versionen nicht voneinander unterschieden.

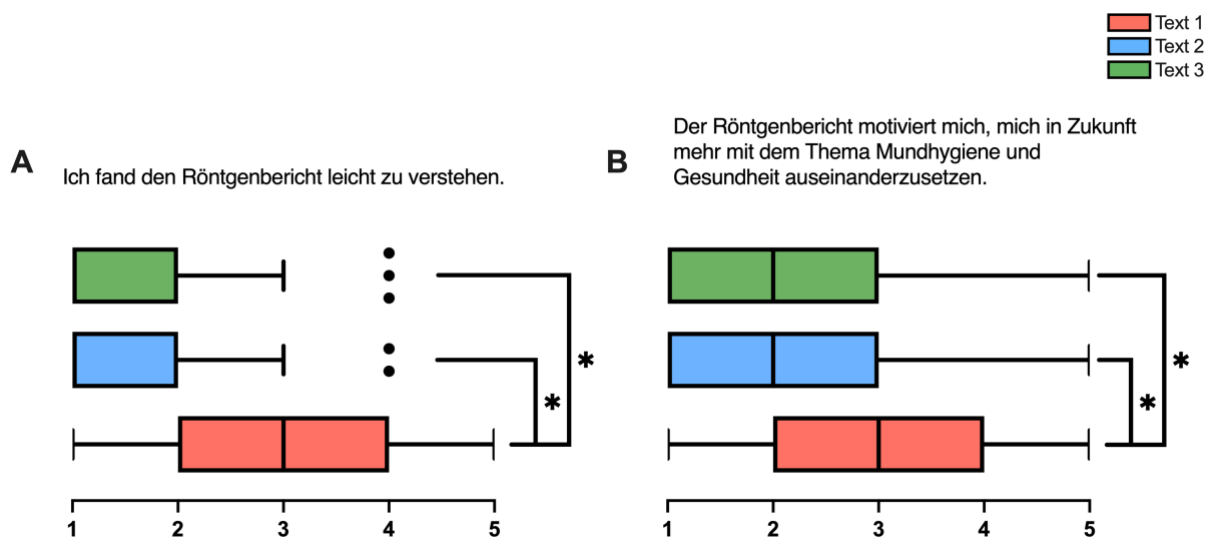


Abbildung 17: Bewertungen des Patientenverständnisses der Röntgenbefunde für die originalen KI-generierten Röntgenberichte (Text 1 – rot), die vereinfachte Version (Text 2 – blau) sowie die stark vereinfachte Version (Text 3 – grün). Evaluiert wurden, ob der Röntgenbericht leicht zu verstehen sei (A) sowie die Motivation sich zukünftig mehr mit Mundhygiene und Gesundheit auseinanderzusetzen (B). Stichprobengröße: $n=50$; $* = p < 0,05$.

3.3.5 Zukünftige Perspektive der Röntgenberichte

Die Frage, ob die Patienten sich wünschen, dass alle ihre medizinischen Berichte so verständlich wie dieser Röntgenbericht formuliert sind, wies ebenfalls signifikante Unterschiede zwischen den Gruppen auf ($F(2,147) = 18,79, p < 0,0001$; Abbildung 18). Text 1 wurde signifikant schlechter bewertet als die Texte 2 und 3, ohne signifikanten Unterschied zwischen den beiden vereinfachten Versionen.

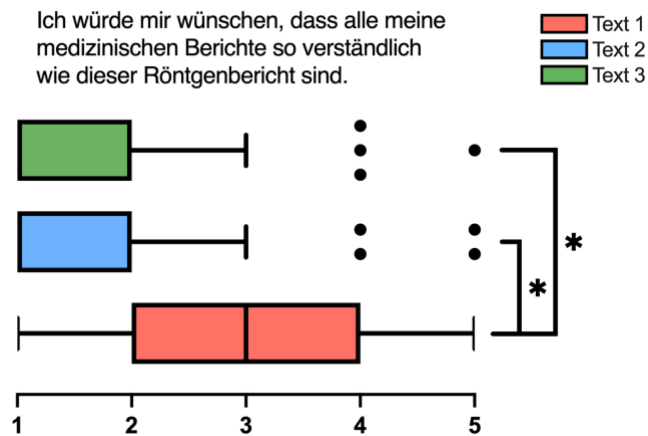


Abbildung 18: Bewertungen des Patientenverständnisses der Röntgenbefunde für die originalen KI-generierten Röntgenberichte (Text 1 – rot), die vereinfachte Version (Text 2 – blau) sowie die stark vereinfachte Version (Text 3 – grün). Evaluiert wurde, ob alle medizinischen Berichte auf diesem Verständnisniveau geschrieben sein sollten. Stichprobengröße: $n=50$; * = $p < 0,05$.

4 Diskussion

Seitdem ChatGPT-3.5 von OpenAI im November 2022 der Öffentlichkeit zur kostenfreien Nutzung zur Verfügung gestellt wurde, übernimmt die Künstliche Intelligenz jeden Tag eine große Bandbreite an Aufgaben, die ihr durch die Eingabe eines einfachen Prompts vom Nutzer gestellt wird. Dabei ist es egal, ob diese pragmatische, analytische oder kreative Natur besitzen. Der Chatbot produziert in vielerlei Hinsicht Ergebnisse, die der menschlichen Sprache äußerst ähnlich sind.

Der öffentliche Zugang ermöglicht prinzipiell jeder Person mit einem Internetzugang, KI-Anwendungen zu nutzen, was einen bedeutungsvollen Schritt in der Demokratisierung einer solchen technologischen Ressource darstellt. Was bisher für viele Menschen als unbekannte „Informatik-Spielerei“ galt, wurde durch die jüngsten Entwicklungen zu einem einfach zu bedienenden Multifunktionswerkzeug, welches in vielen Bereichen des täglichen Lebens Unterstützung bietet, ohne dass der Nutzer über tiefgehende IT-Kenntnisse verfügen muss.

Darüber hinaus hat der freie Zugang zu ChatGPT dazu beigetragen, das Bewusstsein und Verständnis für KI-Technologien in der breiten Bevölkerung zu fördern. Hierdurch werden nicht nur neue Möglichkeiten für Bildung und Forschung geschaffen, sondern auch die Art und Weise zu arbeiten und zu kommunizieren revolutioniert. Bereits heute können die KI-Tools Aufgaben in der Kommunikation, wie beispielsweise dem Verfassen von E-Mails, übernehmen und dem Nutzer so eine bedeutende Arbeitserleichterung verschaffen. Für den medizinischen Bereich lässt sich dies übertragen: Einfache Dokumentationsaufgaben, die an deutschen Kliniken als Zeitfresser gelten und somit personelle Ressourcen beanspruchen [2], könnten möglicherweise zukünftig von Large Language Models teilweise übernommen werden [15, 52].

Ziel dieser Arbeit war es zu untersuchen, ob KI und insbesondere Large Language Models wie ChatGPT in der Lage sind, fachlich und sprachlich korrekte zahnmedizinische Röntgenberichte zuverlässig zu erstellen. Hierauf aufbauend lässt sich perspektivisch auf das Potential zur Effizienzsteigerung und Qualitätsverbesserung in der medizinischen Dokumentation und Kommunikation rückschließen.

Vor dem Hintergrund, dass eine sprachbasierte KI in den klinischen Alltag integriert werden könnte, um medizinische Fachkräfte bei der Befundung und Dokumentation zu entlasten, wurde die Genauigkeit von ChatGPT-4 im Vergleich zu Zahnmedizinierenden bei der Erstellung von Röntgenbefunden untersucht. Anschließend erfolgte die Evaluation, inwiefern Abwandlungen im formulierten Prompt das Ergebnis der KI-generierten Röntgenberichte im Sinne des Prompt-Engineering beeinflussen. Auf diese Weise wurden von ChatGPT vereinfachte Versionen der Röntgenberichte für Patienten unterschiedlicher Bildungsniveaus und ohne medizinische Ausbildung generiert. Abschließend erfolgte die Bewertung der

klinischen Relevanz dieser vereinfachten Versionen durch eine Patientenbefragung, um zu überprüfen, wie Lesbarkeit, Verständlichkeit, Empathie und Textklang, sowie Motivation der verschiedenen Textversionen von den Patienten wahrgenommen werden und welche Art von medizinischen Informationstexten bevorzugt werden.

4.1 Effizienzsteigerung durch KI-gestützte Dokumentation

Wie bereits eingangs dargestellt, nehmen bürokratische Aufgaben wie die Dokumentation im klinischen Alltag durchschnittlich etwa vier Stunden der täglichen Arbeitszeit ein und beanspruchen somit nicht nur personelle und finanzielle Ressourcen, sondern erzeugen auch bei den Behandelnden Unzufriedenheit und Stress [2]. In der vorliegenden Untersuchung konnte am Beispiel der Formulierung eines Röntgenberichts gezeigt werden, dass Studierende der Zahnmedizin bei der Ausformulierung im Fließtext die vollen 30 Minuten benötigten, die ihnen zur Beschreibung eines OPGs zur Verfügung gestellt worden sind. Dem gegenübergestellt wurde ein zeitliches Limit von 10 Minuten für das Ausfüllen einer Checkboxliste. Basierend auf dieser Checkboxliste generierte die KI innerhalb weniger Sekunden eigenständig einen Fließtext. Die Tatsache, dass alle Studierenden die gesamte zur Verfügung stehende Zeit von einer halben Stunde voll ausnutzen, zeigt auch in diesem Fall, wie zeitintensiv die Dokumentation eines Röntgenbefundes sein kann. Konträr dazu konnte durch das Studiendesign gezeigt werden, dass der Einsatz von KI die Effizienz der Dokumentation steigern konnte und 20 Minuten pro Bericht eingespart wurden. Die Studie zeigt somit, dass KI-generierte Berichte eine signifikante Reduktion der Zeit und des Aufwands für die Erstellung medizinischer Dokumentation ermöglichen.

Dies wurde bereits von Baker et al. beschrieben, die in ihrer Studie „ChatGPT's Ability to Assist with Clinical Documentation: A Randomized Controlled Trial“ zeigten, dass die Generierung einer Krankengeschichte durch ChatGPT schneller erfolgte als das Abtippen dieser durch den Behandler. Weiterhin erstellte die KI im Vergleich zu anderen Methoden, wie beispielsweise dem Diktieren, eine detailliertere und strukturiertere Dokumentation [55].

Der Einsatz einer Checkboxliste könnte darüber hinaus auch als Gedankenstütze und Kontrolle der Vollständigkeit dienen, um Flüchtigkeitsfehler im durchaus stressigen Berufsalltag zu minimieren. So konnte in einer Studie von Storesund et al. gezeigt werden, dass durch den Einsatz von prä- und postoperativen SURPASS-Checklisten sowie intraoperativen WHO-Checklisten die Anzahl an Komplikationen, Notfall-Reoperationen und ungeplanten 30-Tage-Wiedereinweisungen reduziert werden konnten [83]. Eine Limitation dieser Studie war jedoch die Gestaltung der Checkboxliste, der es an Übersichtlichkeit mangelte. Für zukünftige Studien sollte diese effizienter gestaltet werden. Eine Darstellung der einzelnen Zähne im etablierten Zahnschema könnte das Ausfüllen der Checkboxliste

erleichtern und weniger anspruchsvoll gestalten. Aufgrund des verwendeten Tabellenformats glich die Zuordnung der Zähne in der vorliegenden Studie eher einer Konzentrationsübung.

Abbildung 19 zeigt ein Beispiel eines Zahnschemas, bei der die Darstellung der natürlichen Anordnung folgt und somit die Übertragung von Befunden erleichtert wird. Nichtsdestotrotz wurden Pathologien auf den Röntgenbildern, die über eine Checkliste abgefragt wurden, weniger häufig übersehen.

4.2 Vergleich der Textqualität und Lesbarkeit

Ein weiterer Aspekt, der zur Bewertung der inhaltlichen Qualität herangezogen werden muss, ist die Anzahl der in den Röntgenberichten festgestellten Fehlern. Im Rahmen der vorliegenden Studie konnte gezeigt werden, dass Studierende im Vergleich zu ChatGPT signifikant mehr Rechtschreibfehler, Interpunktionsfehler und Grammatikfehler in ihren Röntgenberichten machten, während die KI alle Texte frei von sprachlichen Fehlern generierte. Der hohe Fehlerquotient der Studierenden könnte durch das festgelegte Zeitlimit bedingt sein, das jedoch gleichzeitig den hohen Zeitdruck im späteren Berufsalltag simulierte.

Weitere Metriken mit denen die Lesbarkeit der Texte beschrieben werden kann sind zum einen der LIX und zum anderen der FRE-Score. Diese werden in der Literatur oftmals zur Quantifizierung der akademischen Anforderungen an den Leser genutzt. Dazu wird unter anderem die durchschnittliche Silbenzahl pro Wort sowie die durchschnittliche Länge eines Satzes bestimmt, um die Komplexität eines Textes zu beschreiben. Ein niedriger FRE-Score lässt sich mit einem wissenschaftlichen Schreibstil assoziieren, da dieser oft durch den Gebrauch längerer Wörter und komplexerer Satzstrukturen charakterisiert ist. Im Gegensatz dazu deutet ein hoher FRE-Score auf eine einfachere Sprache mit kürzeren Sätzen hin, die ein geringeres Bildungsniveau zum Verständnis erfordert. Der LIX-Score weist aufgrund seiner Berechnung gegensätzliche Werte zum FRE-Score auf, weshalb der LIX-Score mit zunehmender sprachlicher Komplexität ansteigt. In der vorliegenden Studie konnte gezeigt werden, dass der FRE-Score kaum signifikante Unterschiede in der Lesbarkeit zwischen den KI-generierten Röntgenberichten und den von Studierenden verfassten Berichten zeigte. Dieses Ergebnis deutet darauf hin, dass die KI ihre Texte auf demselben akademischen Niveau formulierte, wie die Studierenden. Für die klinische Praxis bietet dieses Resultat hinsichtlich der Anwendbarkeit von generierten Texten ein großes Potential.

Darüber hinaus stellt die Abwesenheit von Rechtschreib-, Grammatik- und Interpunktionsfehlern in den von der KI generierten Texten einen erheblichen Vorteil dar, da sie die Lesbarkeit und Verständlichkeit erheblich verbessern. Ein fehlerfreier Text wird außerdem im Rahmen einer professionellen Kommunikation vorausgesetzt. Im Gegensatz

dazu weisen die von den Studierenden erstellte Texte häufige Flüchtigkeitsfehler und sprachliche Fehlformulierungen auf. Diese Fehler könnten in der praktischen Anwendung potenziell zu Missverständnissen führen und die Kommunikation zwischen medizinischen Fachkräften und Patienten beeinträchtigen. Eine beispielhafte Formulierung eines Studierenden macht dies deutlich: *„Es lässt sich jedoch unschwer eine Extraktionsalveole oder Ähnliches vermuten, da der Knochen nicht gut erkennbar ist.“*. Der Begriff „unschwer“ wurde hier so eingesetzt, dass die Bedeutung des Satzes nicht klar hervorgeht. Ein weiteres Beispiel zeigt eine Beschreibung eines Studierenden, der wahrscheinlich Implantate in den Frontzahnregionen beschreiben wollte, jedoch formulierte: *„schraubenförmige Gebilde an den Zähnen 12, 11, 21, 22“*. Auch aus diesem Satz geht keine klare Bedeutung für den Röntgenbericht hervor. Formulierungen wie diese lassen sehr viel Interpretationsspielraum, da sowohl Zahnimplantate als auch Schrauben oder Pins im Rahmen einer knöchernen Augmentation gemeint sein können. Es können sogar in situ befindliche Osteosyntheseschrauben nach einer Frakturversorgung nicht ausgeschlossen werden.

Dem gegenübergestellt zeigt die Anwendung der KI erhebliches Potenzial, die medizinische Kommunikation klarer und fehlerfreier zu gestalten. Dies könnte zu einer erheblichen Reduzierung von Missverständnissen und somit zu einer Verbesserung der Verständlichkeit der medizinischen Berichterstattung führen, was letztlich sowohl die Effizienz der medizinischen Fachkräfte als auch die Qualität der Patientenversorgung erhöhen kann.

4.3 Informationsgenauigkeit und Vollständigkeit

Zur Evaluation der Informationsdichte und Vollständigkeit wurde die Anzahl der durch die Studierenden angekreuzten Kästchen in der Checkboxliste mit der Anzahl der hieraus in die KI-generierten Röntgenberichte übernommenen Diagnosen verglichen. Dabei wurde festgestellt, dass aus der zweiten und dritten Tabelle der Checkboxliste signifikant weniger Aussagen übernommen wurden und somit weniger relevante Informationen in die Berichte inkorporiert worden sind. In der Veröffentlichung „Differentiating ChatGPT-Generated and Human-Written Medical Texts: Quantative Study“ zeigten Liao et al., dass KI-generierte Texte zwar eine gute Lesbarkeit und grammatikalische und sprachliche Korrektheit, menschlich formulierte Texte allerdings ein höheres Maß an Konkretheit und medizinischer Informationsdichte aufwiesen [54]. Ähnliches lässt sich auch in der vorliegenden Studie zeigen, wobei die Gründe hier jedoch in der Formatübertragung und Übersetzung der Excel-Datei liegen könnten.

4.3.1 Einfluss des Formatierungsdesigns auf die Informationsdichte von KI-generierten Berichten

Während der Vorbereitung dieser Studie wurden verschiedene Versuche unternommen, mit ChatGPT Texte auf Basis der Checkboxlisten zu generieren, die je nach Format der hochgeladenen Datei unterschiedliche Ergebnisse lieferten. Es wurden mehrere Formate getestet, darunter Word-Dokumente, CSV-Dateien und Excel-Dateien, wobei letztere sich schließlich am besten zur Translation eigneten. Zunächst wurde versucht, alle Aussagen horizontal und dann vertikal auf einer Excel-Seite zu platzieren. Dabei trat das Problem auf, dass ChatGPT ab einer bestimmten Zeilenzahl die Datei nicht mehr vollständig lesen konnte, was zu einem Abbruch der Analyse und zufälligen Kombinationen von Spaltenüberschriften mit den darunterliegenden Informationen führte. Eine weitere Herausforderung bestand darin, dass die Nummerierung der Zähne, die gemäß dem europäischen Zahnschema der Fédération Dentaire Internationale (FDI) beschrieben werden, von der KI fehlerhaft interpretiert wurden. Die Tabellenblätter, die Angaben zu den einzelnen Zähnen enthielten, zeigten dementsprechend eine geringere Wahrscheinlichkeit zur Übernahme der Aussagen in den Bericht. Dies zeigte sich auch in der Tatsache, dass das erste Tabellenblatt, das anders als das zweite und dritte Tabellenblatt keine Informationen zu einzelnen Zähne enthielt, eine deutlich geringere Fehleranfälligkeit und eine höhere Übernahmechance der Aussagen aufwies. Einige der auftretenden Probleme konnten schließlich gelöst werden, indem jede Überschrift einem individuellen Tabellenblatt zugewiesen wurde, das jeweils den Titel der entsprechenden Überschrift aus der Checkboxliste erhielt. Die Spalte, die das „X“ zur Markierung der Aussagen enthielt, wurde als „Checkbox“ bezeichnet. Es wurden dann die Zahn-Nummerierungen von 18 bis 48 vertikal untereinander aufgelistet. Die dazugehörigen beschreibenden Aussagen wie beispielsweise die Bewertung einer Krone als „suffizient“ oder „insuffizient“ wurden darunter aufgeführt.

Insgesamt ergaben sich bei dieser Darstellungsform 21 Tabellenblätter, die von der KI analysiert werden sollten. Ein besonderes Augenmerk wurde auf das Prompt-Design gelegt, das entscheidende Informationen zur korrekten Interpretation der erstellten Datei durch die KI enthalten musste. Es wurde sichergestellt, dass klar definiert war, in welcher Spalte das „X“ platziert war, das als positive Aussage in den Bericht aufgenommen werden sollte.

„Bitte erwähne im Röntgenbericht nur die Aussagen, bei denen ein Kästchen mit einem X markiert ist. Die nicht mit X markierten Aussagen sollen nicht im Bericht aufgenommen werden.“

Auch wenn die schlussendlichen Ergebnisse eine hohe Dichte an übernommenen Aussagen aufwiesen, zeigte sich hierbei, dass das Format in erheblichen Maß die Informationen beeinflusste, die durch die KI übernommen wurden. Dennoch verdeutlicht die Untersuchung,

wie effizient das Prompt-Design die Ausgabertexte prägte. Weitere Verbesserungen des Prompts führen zu einem präziseren Ergebnis, was vor allem in zukünftigen Studien Beachtung finden sollte. Beispielsweise könnten hier weitere Tabellenblätter hinzugefügt werden, sodass jeweils weniger Informationen enthalten sind, die analysiert werden müssen. Ein weiterer Ansatz könnte in der Verwendung von dem in Abbildung 19 aufgezeigten Zahnschema liegen, dass die KI gegebenenfalls mit weiterem Voranschreiten der Entwicklung besser analysieren kann.

4.3.2 Analyse und Korrektur von Fehlinterpretationen in KI-generierten Berichte

Die Aussagen, die durch ChatGPT ausgelassen worden sind, wurden in einem weiteren Arbeitsschritt gemeinsam mit den Fehlinterpretationen und -formulierungen im Sinne von Referenztexten manuell verbessert und ergänzt. Anschließend wurden diese Referenztexte mit den KI-generierten Texten unter Anwendung des BERT-Scores, einem anerkannten Maß für die Textähnlichkeit bei maschinell erzeugten Übersetzungen, Zusammenfassungen und Texten, verglichen [78]. Die Precision (P) misst hierbei den Anteil der Token, die „präzise“ den Originaltext und deren Bedeutung wiedergeben. Diese zeigt in der Studie genau wie der Recall (R), welcher ein Maß für die Informationsgenauigkeit der Texte verglichen mit den Referenztexten darstellt, einen Wert von 0.965 (P) und 0.96 (R). Der F1-Score bildet schließlich das harmonische Mittel dieser Komponenten. Wenn nun zwei identische Texte gegenübergestellt werden, dann beträgt der F1-Score den Wert 1.00. Gegensätzliches kann gezeigt werden, wenn ein Text mit einem Referenztext ohne thematische Überschneidung verglichen wird. Beispielsweise erreicht ein Wikipedia-Artikel über das Atom und ein Auszug aus einem Märchen der Gebrüder Grimm beim BERT-Score einen F1-Wert von 0,12.

Die Ergebnisse der Studie belegen, dass sowohl die inhaltlichen Erweiterungen als auch die sprachlichen Umformulierungen durch die KI nur minimal waren, wobei eine hohe Ähnlichkeit zu den ursprünglichen Referenztexten bestand. Dennoch zeigten sich einige wiederkehrende Fehlinterpretationen durch die KI, welche zu einer Verbesserungsbedürftigkeit der Texte führte. Ein Beispiel hierfür ist das Tabellenblatt „*Bewertung des Knochens*“, welches Informationen darüber liefern sollte, ob ein generalisierter Knochenabbau im Unter- und Oberkiefer besteht. Weiterhin sollten die vertikalen Knocheneinbrüche in den jeweiligen Zahnpositionen beschrieben werden. Folgendes wurde in einem Fall durch die KI generiert:

„Es gibt einen Knochenabbau im Unterkiefer sowie im Oberkiefer, besonders in den Regionen der Zähne 38, 37, 36, 46, 47 und 48.“

Die Information über die Richtung des Knochenabbaus (vertikal und / oder horizontal), wurde hier ausgelassen. In einem anderen Beispiel wurde der Knochenabbau garnicht erwähnt, vielmehr wurde der Knochen als „auffällig“ beschrieben, was jedoch nicht die Informationen, die die KI durch die vorgelegte Excel-Datei erhalten hat, widerspiegelte:

„Beim Knochen gibt es Auffälligkeiten sowohl im Unterkiefer als auch im Oberkiefer, speziell in den Bereichen 26, 37 und 36.“

Auch in diesem Fall wurde die Aussage im Referenztext korrigiert zu: „Die Bewertung des Knochens zeigt im Unterkiefer einen generalisierten horizontalen Knochenabbau, insbesondere im Bereich der Zähne 36 und 46 zeigen sich vertikale Knocheneinbrüche“.

Darüber hinaus wurde das Tabellenblatt „Sapientes“ häufig ausgelassen, was auf die komplexere Darstellung der Informationen in drei, anstatt zwei Spalten zurückzuführen sein könnte. Es wurden auch häufig Formulierungen wie *„Implantate sind bei den Zähnen 37, 36 und 34 eingesetzt“* verbessert, da diese in der Praxis angewandt zu Missverständnissen führen könnte.

4.4 Patientenperspektive und -vertrauen

Die Verständlichkeit von Informationsschriften für Patienten hängt stark von deren Bildungsstand ab. Patienten mit niedrigem Bildungsniveau haben oft Schwierigkeiten, medizinische Texte zu verstehen, was ihre Fähigkeit beeinträchtigen kann, fundierte Entscheidungen über ihre Behandlung zu treffen [61]. Der Einsatz von KI könnte hierbei helfen, indem durch sie patientenfreundliche Texte erstellt werden, die komplexe Informationen verständlicher machen, ohne den Inhalt zu verändern. Zusätzlich kann ein Text, welcher der ärztlichen Kommunikation gedient hat und für den medizinisch ungebildeten Patienten unverständlich ist, mithilfe von KI umformuliert werden, sodass er für den Patienten besser verständlich ist. Auf der Grundlage eines Textes kann so theoretisch durch die Anwendung unterschiedlich ausformulierter Prompts ein breites Bildungsspektrum adressiert werden.

Die Auswertung der patientenseitigen Evaluationen zeigte signifikante Unterschiede in der Bewertung der verschiedenen Textversionen. So war der erste Prompt so formuliert, dass der Röntgenbericht aus der Perspektive des behandelnden Zahnarztes geschrieben wurde (Text 1). Im weiteren Verlauf wurde das Ergebnis mit einer Variante des Prompts durch die KI so umformuliert, dass der Text für Patienten verständlicher sein sollte (Text 2). Der dritte Prompt zielte darauf ab, den Text weiter zu simplifizieren, sodass alle Patienten unabhängig ihres Bildungsniveaus den Röntgenbericht verstehen würden (Text 3).

4.4.1 Verständlichkeit und Informationsgehalt der KI-generierten Berichte

Der den Patienten vorgelegte Text 1 wies einen FRE-Score von 51 auf, während im Vergleich Text 3, also der durch die KI am stärksten vereinfachte Bericht, einen FRE-Score von 56 ergab. Die Anwendbarkeit dieses Scores lässt sich auch klinisch verifizieren, indem die Patienten die Aussage *„Ich fand den Röntgenbericht leicht zu verstehen“* für die beiden vereinfachten Röntgenberichte (Text 2 und 3) als zutreffend bewerteten, während sie Text 1 als weniger

verständlich einstufen. Ein möglicher Erklärungsansatz hierfür ist, dass in den überarbeiteten Texten Fachbegriffe entweder erläutert oder seltener verwendet wurden. Ein Beispiel hierfür ist die Ersetzung des Begriffs „Implantat“ durch die verständlichere Formulierung „künstliche Zahnwurzel“. Auch die Aussage ‚Die verwendeten Fachbegriffe wurden klar erklärt‘ wurde – analog zur ersten Frage – mit signifikanten Unterschieden zwischen Text 2 und 3 im Vergleich zu Text 1 bewertet. Dies spiegelt sich auch in den Ergebnissen von Frage 3 wider, die untersuchte, inwieweit der Befund ohne zusätzliche Unterstützung verstanden werden konnte. Auch hier schnitt Text 1 signifikant schlechter ab als die beiden überarbeiteten Röntgenberichte. Die Ergebnisse belegten, dass die KI in der Lage ist, Formulierungen gezielt an die Anforderungen des jeweiligen Prompts anzupassen. Zudem wird die Kompetenz der KI bei der Erklärung medizinischer Fachbegriffe deutlich, was im klinischen Alltag zu einer verbesserten Verständlichkeit medizinischer Texte beitragen kann – mit potenziell positiven Effekten auf das Vertrauen der Patienten sowie deren Behandlungssadhärenz.

Die Aussage *„Die Informationen im Röntgenbericht waren ausreichend detailliert, um meine Fragen zu beantworten“* wurde von den Patienten für Text 2 und 3 als zutreffend bewertet, während Text 1 deutlich schlechter abschnitt. Diese Bewertung lässt sich mit den bereits genannten ausführlichen Erklärungen medizinischer Fachbegriffe in Text 2 und 3 in Verbindung bringen. Auch die Aussagen *„Der Röntgenbericht hilft mir dabei, sinnvolle Fragen an meinen Behandler zu stellen“* sowie *„Nach dem Lesen des Röntgenberichts kann ich meine Diagnose mit dem Behandler besprechen und ein Gespräch auf Augenhöhe führen“* wurden für Text 2 und 3 überwiegend positiv bewertet, während Text 1 erneut signifikant schlechter beurteilt wurde. Diese Ergebnisse deuten darauf hin, dass die Fachsprache und die komplexe Beschreibung anatomischer Strukturen in Text 1 für Laien schwer verständlich waren und somit keine ausreichende Grundlage für Rückfragen oder ein informatives Arzt-Patienten-Gespräch boten. Ein zentraler Anspruch medizinischer Kommunikation sollte jedoch die Verständlichkeit und Interaktivität sein, um ein vertrauensvolles Verhältnis und eine partnerschaftliche Entscheidungsfindung zu ermöglichen. Die Bewertungen zu Text 1 legen nahe, dass dies aufgrund der wissenschaftlich geprägten Sprache von den Patienten nicht als gegeben empfunden wurde.

Eine klarere und einfachere Formulierung kann dazu beitragen, dass Patienten besser informiert und dadurch in der Lage sind, relevante Fragen zu stellen und fundierte Diskussionen mit ihren Behandlern zu führen. Dies verbessert nicht nur die Patientenzufriedenheit, sondern kann auch die Qualität der Behandlung durch eine verbesserte Kommunikation, ein gestärktes Arzt-Patient-Verhältnis und ein besseres Verständnis der Diagnose und der Behandlungsmöglichkeiten fördern. Der Einsatz von KI könnte so im klinischen Kontext dafür sorgen, dass Patienten nicht auf Informationen aus dem

Internet, die häufig Fehler enthalten, zurückgreifen müssen, sondern mit ihren Bedenken und Fragen direkt ihren Behandler konsultieren [63].

4.4.2 Empathie und Ton

Um die Situation des Patienten und die durch die Patienten subjektiv wahrgenommene Empathie zu bewerten, wurden zwei weitere Fragen in den Bewertungsbogen aufgenommen. Zum einen sollten die Patienten folgende Aussage bewerten: *„Der Röntgenbericht vermittelte eine empathische Haltung gegenüber meinem Verständnisniveau.“*. Die Ergebnisse zeigen, dass die Patienten Text 2 und 3 signifikant häufiger als empathischer empfanden als Text 1. Auch der durch die Formulierungen vermittelte Respekt sollte durch folgende Aussage bewertet werden: *„Der Ton des Röntgenberichts war respektvoll und berücksichtigte meine Situation als Patient.“*. Auch hier wurden die modifizierten Texte signifikant besser bewertet als der ursprüngliche Röntgenbericht. Das könnte unter anderem an den subjektiv wahrgenommenen respektvollen Formulierungen der KI gelegen haben. Ein Beispiel hierfür zeigt die Empfehlung zu einer Zystektomie, die als *„kleine Operation“* beschrieben wird und so die Befürchtung des Patienten vor einem umfangreicheren Eingriff und damit verbundenen Komplikationen reduzieren könnte. Die KI generierte dazu folgenden Satz:

„Abhängig von den Ergebnissen könnten wir eine kleine Operation in Betracht ziehen, um das Problem mit der Zyste zu lösen, sowie weitere Untersuchungen, um sicherzustellen, dass alles in Ordnung ist.“

Auch die Aussage, dass der Behandler *„sichergehen will, dass alles in Ordnung ist“* wurde von den Patienten wahrscheinlich als empathische Haltung aufgenommen, während aus der Perspektive des Zahnarztes dazu geschrieben wurde:

„Es sind zystische Veränderungen im linken Ramus vorhanden, die sowohl scharf als auch unscharf begrenzt sind, jedoch ohne Kontinuitätsunterbrechung der Compacta. Aufgrund der Befunde empfehle ich eine weitere 3D-Bildgebung, etwa durch DVT, MRT oder CT.“

Diese Formulierung ist zwar medizinisch korrekt, dennoch werden hier die persönlichen Befürchtungen und Sorgen des Patienten nicht adressiert. Dazu kommen Formulierungen wie *„Wenn Sie Fragen haben oder weitere Erläuterungen benötigen, stehen wir Ihnen gerne zur Verfügung.“*, welche dem Patienten signalisieren, dass der Behandler für Rückfragen offen ist und bereitsteht, diese zu beantworten.

Konträr zur Empathie, die über die Informationsschrift transportiert werden soll, stellt die Abschwächung von Formulierungen ein Risiko für Fehlinterpretationen durch den Patienten dar. So kann beispielsweise eine *„kleine Operation“* ein Eingriff unter Lokalanästhesie bedeuten. Bei größeren Eingriffen, welche teilweise nötig sind, um Zysten zu entfernen, würde

der Patient von dem Umfang des Eingriffes wahrscheinlich überrascht werden. So liegt die Verantwortung der Formulierungskontrolle letztlich wieder beim Behandler.

Zusammenfassend konnte hier deutlich gezeigt werden, dass KI-Anwendungen dazu in der Lage sind, empathisch klingende medizinische Informationsschriften zu erstellen. Dies könnte auch dabei helfen, das Vertrauen der Patienten in KI zu stärken, um den Einsatz dieser im klinischen Kontext zu erweitern. Bisher konnte in einer Studie von Juravle et al. gezeigt werden, dass Patienten den Ergebnissen der künstlichen Intelligenz weniger Vertrauen entgegenbringen als denen der menschlichen Behandler, obwohl die Patienten in dieser Studie darüber informiert worden sind, dass die Technologie oftmals bessere Ergebnisse produzierte als die menschlichen Behandler [68].

Im Gegensatz dazu konnte in der vorliegenden Studie gezeigt werden, dass Patienten die Formulierungen des ihnen vorgelegten KI-generierten Röntgenberichts gegenüber konventionellen Informationsschriften präferieren würden. Dies konnte in der Analyse folgender Aussage gezeigt werden: *„Ich würde mir wünschen, dass alle meine medizinischen Berichte so verständlich wie dieser Röntgenbericht sind.“*. Diese Aussage wurde ebenfalls für Text 2 und 3 zu einem großen Teil als zutreffend bewertet. Das Ergebnis unterstreicht den Wunsch der Patienten nach einer respektvollen Kommunikation auf Augenhöhe sowie nach verständlichen Erklärungen medizinischer Fachbegriffe – unabhängig davon, wer die Informationsschrift verfasst hat. Auch die Rückmeldungen der Teilnehmenden spiegeln dies wider: Viele äußerten sich positiv überrascht über die Fähigkeiten der Künstlichen Intelligenz und wünschten sich, dass künftig generell in dieser verständlichen und zugewandten Weise im medizinischen Kontext mit ihnen kommuniziert wird.

Da das Studiendesign nur die Bewertung von drei Röntgenberichten vorsah, sollte hier angesetzt werden, um weiter an verschiedenen Befehlen zu forschen, da das Prompt-Design einen erheblichen Einfluss auf die Qualität der KI-generierten Ausgabe zeigte. Hier könnten unterschiedliche Perspektiven oder Adressaten im Prompt inkorporiert werden, um diese im Patientenkontext weiter zu erproben. Da sich die vorliegende Studie nur auf Röntgenberichte bezieht, könnten hier weitere medizinische Informationsschriften für den Patienten neu generiert werden. Ein Beispiel könnte hier ein Pathologiebefund sein, welcher für den Patienten verständlicher umformuliert werden könnte. Zusätzlich sollte eine inhaltliche Validierung im Fokus stehen, sodass jene Texte, die durch die KI generiert werden, auch professionell bewertet werden. Abschließend könnten inhaltliche Abfragen das objektive Verständnis des Patienten evaluieren, da die vorliegende Studie nur auf das subjektive Patientenverständnis beschränkt war.

4.5 Sicherheitsstrukturen in der Zusammenarbeit mit KI

Ein weiterer Punkt, der das Patientenvertrauen im Umgang mit KI im klinischen Setting erhöht ist das Etablieren von robusten Sicherheitsstrukturen. Ein Rahmen hierfür bietet die europäische Datenschutzgrundverordnung (DSGVO) und die ärztliche Schweigepflicht, welche das Vertrauensverhältnis zwischen Patient und Behandler wahren soll [84]. Prinzipiell soll ein Patient auf Grundlage der offengelegten Daten nicht identifiziert werden können. Dies gestaltet sich im medizinischen Rahmen allerdings aufgrund der Individualität der Gesundheitszustände schwer und stellt eine ethische Herausforderung dar. Welche Informationen für die Identifikation einer Person notwendig sind ist jedoch umstritten [84]. Darüber hinaus soll die Diskriminierung durch KI durch den Gleichheitssatz auf europäischer Ebene durch die Artikel 20-26 der GRC (Charta der Grundrechte der Europäischen Union) reguliert werden. Diese Artikel sollen die Gleichheit aller Bürger und die Nichtdiskriminierung durch Merkmale wie Ethnie, Religion, Sprache, Geschlecht, Alter oder Behinderung garantieren [85]. Dazu muss im Kontext medizinischer Anwendung die Qualität der Quellen, die als Grundlage für die Trainingsdaten der KI gewählt werden, strengen Kontrollen unterzogen werden, um im Algorithmus enthaltene Fehler wie beispielsweise die Diskriminierung verschiedener Geschlechter, Ethnien, Sexualitäten oder Altersgruppen zu minimieren.

Für die klinische Integration der KI ist es also notwendig, dass die Systeme regelmäßig strengen Überprüfungen unterzogen werden und ein effektives Qualitätsmanagement implementiert wird, um mögliche Fehler zu vermeiden und schnellstmöglich zu beheben. Um die „Black-Box Problematik“ zu adressieren, sollte auch hier eine genügende Überprüfbarkeit von Algorithmen vorherrschen. Es müssen Mechanismen geschaffen werden, die es ermöglichen, die Entscheidungspfade von Algorithmen nachzuvollziehen, sodass diese der menschlichen Kontrolle unterzogen werden können. Schließlich ist es aufgrund des Patientenrechtegesetz obligat, den Patienten über die Nutzung von KI aufzuklären, um auch die Entscheidung gegen die Verwendung dessen in die Behandlung miteinzubeziehen [1, 86]. Hierbei ist es wichtig, eine suffiziente Aufklärung zu leisten und den Patienten über Vorteile und potenzielle Risiken aufzuklären. Das Einholen der informierten Einwilligung der Patienten sollte bei der Verwendung von KI zu einem Standardverfahren etabliert werden. Darüber hinaus könnten Patienten aktiv in einen Feedbackprozess eingebunden werden, um ähnlich wie in dem vorliegenden Studiendesign ihre Erfahrungen mit KI-gestützten Behandlungen weiter voranzutreiben. Dies verbessert nicht nur die Technologie selbst, sondern könnte auch das Vertrauen in solche Systeme weiter stärken.

Während der Untersuchung der KI-generierten Röntgenberichte sind die Sicherheitsstrukturen, die durch ChatGPT geboten werden in einem Punkt besonders aufgefallen. Durch das Hinzufügen von Warnhinweisen wie *„Dieser Bericht dient nur zu*

Informationszwecken und ersetzt nicht die fachliche Meinung eines qualifizierten medizinischen Fachmanns. Bei weiteren Fragen oder Bedenken wird empfohlen, einen Spezialisten zu konsultieren.“ stellt OpenAI sicher, dass die Nutzer von ChatGPT daran erinnert werden, dass KI-generierte Inhalte nicht die professionelle medizinische Beratung ersetzen können. Darüber hinaus tragen diese Formulierungen dazu bei, die Grenzen der KI zu verdeutlichen, indem sie die Notwendigkeit einer menschlichen Überprüfung und Bestätigung betonen.

Schließlich geht aus der Studie auch hervor, dass die KI-generierten Texte und Resultate einem hohen Standard entsprechen, sowohl auf sprachlicher als auch auf Inhaltlicher Ebene. Dennoch wurde auch deutlich, dass die Kontrolle durch den Behandler bisher unerlässlich ist, um die Korrektheit der Ergebnisse zu wahren, sodass in der medizinischen Kommunikation keine Fehler und Missverständnisse auftreten. Der Begriff des laut Beck et al. „symbolischen Haftungsknechts“ [12] verdeutlicht in diesem Hinblick, dass der Behandler für die Fehler der KI verantwortlich gemacht würde.

Daraus resultiert, dass eine angepasste medizinische Ausbildung, welche die Anwendung von KI und die Vermittlung eines tiefgehenden Technologieverständnisses beinhaltet, vor der Etablierung solcher Algorithmen eine absolute Notwendigkeit darstellen [87]. An der Universitätsmedizin Mainz wurde zum Wintersemester 2022/23 ein Kurs zum Thema „KI in der Radiologie“ im Rahmen der Wahlpflichtwochen des Medizinstudiums angeboten. Eine Erweiterung solcher Kurse und die Einbindung in den regulären Lehrplan ist ein entscheidender Schritt, der den Einsatz von Künstlicher Intelligenz in das klinische Geschehen in Zukunft ermöglichen könnte und einen möglichst sicheren Umgang mit solchen Technologien gewährleistet. Durch eine umfassende Ausbildung auf diesem Gebiet können die zukünftigen Behandler die Potenziale der KI kennenlernen und durch die Integration die Qualität der Patientenversorgung und deren Sicherheit zukünftig verbessern.

4.6 Perspektiven der Implementierung

Es konnte bereits in einigen medizinischen Fachrichtungen wie beispielsweise der Dermatologie, der Radiologie oder Pathologie gezeigt werden, dass der Einsatz von KI sich als sinnvoll erwiesen hat und eine Alternative zu konventionellen Behandlungsabläufen und Diagnosemöglichkeiten darstellt [5-8]. Was die klinische Dokumentation betrifft, konnten Baker et al. in ihrer Studie zur KI-gestützter Erstellung von Krankengeschichten zeigen, dass der Einsatz von Sprachmodellen wie ChatGPT eine Arbeitsentlastung darstellen kann [55]. In der vorliegenden Studie konnte zudem gezeigt werden, dass Patienten KI-generierte Röntgenberichte den konventionellen medizinischen Informationsschriften bevorzugen würden. Dennoch beruht die Studie auf einer rein experimentellen Umsetzung und ist in den

untersuchten Strukturen und Abläufen noch nicht in den klinischen Ablauf integrierbar. Um die KI-gestützte Dokumentation also für den medizinischen Arbeitsablauf zugänglich zu machen, sollten weitere Ressourcen in die Weiterentwicklung auf diesem Gebiet genutzt werden, sodass sprachbasierte Algorithmen zukünftig optimiert Anwendung finden können. Dazu sind ein gesundheitspolitischer Fokus und die Investition von finanzieller Förderung in die Forschung der angewandten KI notwendig. Darüber hinaus erfordert die Implementierung auch interdisziplinäre Zusammenarbeit zwischen den Feldern der Informatik, Ethik, Medizin und Datenwissenschaft, um Sicherheit und umsetzbare Strukturen zu schaffen, die die Praktikabilität verbessern.

Ein weiterer Ansatz zur Erhöhung der patientenbezogenen Sicherheit ist das Trainieren eines Large Language Models, das rein auf die klinische Dokumentation ausgelegt ist. Dadurch könnten die kontrollierten Patientendaten, auf denen das Training einer solchen KI basiert, selektiert werden, sodass die dem Algorithmus enthaltene Diskriminierung anderer LLM wie ChatGPT limitiert wird [37, 39]. Zur Verarbeitung dieser Daten müssen die Patienten jedoch aufgrund der Datenschutzrichtlinie wie zuvor erwähnt zustimmen und die Daten müssen so verarbeitet werden, dass selbst bei einer Sicherheitslücke kein Rückschluss auf die Identität möglich ist. Zusätzlich beschreibt der mit KI in Zusammenhang stehende Begriff „Big Data“, dass ein solcher Algorithmus auf der Grundlage einer Vielzahl von Daten trainiert wird. So zeigte sich in bisherigen Versuchen, dass die Daten einer einzelnen medizinischen Institution nicht ausreichen. Vielmehr müssten die patientenbezogenen Daten mehrerer Organisationen im Sinne eines Multicenter-Ansatzes zusammengelegt werden, um ein valides Ergebnis zu erzeugen [87]. Im Vergleich zu solchen Modellen wie ChatGPT sollte eine medizinisch angewandte KI modulierbar sein, sodass Fehlerquellen unumgänglich behoben werden können und das Trainieren des Algorithmus adaptiv gestaltet werden kann.

Parallel dazu muss die medizinische Ausbildung angepasst werden, um zukünftige Ärzte auf die Arbeit mit KI-Anwendungen vorzubereiten. Dies beinhaltet die Vermittlung von grundlegendem Technologieverständnis bereits während des Medizinstudiums, insbesondere im Hinblick auf KI-Anwendungen in der Medizin und deren ethische Folgen. In Verbindung dazu haben Studien bereits gezeigt, dass der Einsatz von virtuellen Lernmethoden und modernen Technologien die Lernleistung von Studierenden erheblich verbessert, indem sie die vermittelten Informationen anschaulicher und zugänglicher machen [88].

Ein weiterer wesentlicher Aspekt der Integration von KI in die Medizin ist der Zugang zu den neusten Technologien und der Gleichheit der Behandlungsangebote an die Patienten. Es ist unerlässlich, dass die Demokratisierung der KI-Technologie vorangetrieben wird, um sicherzustellen, dass alle Patienten gleichberechtigten Zugang zu hochwertigen Behandlungen erhalten. Die Entwicklung und Implementierung von KI-Systemen müssen so gestaltet sein, dass sie den Grundsatz der Gleichheit wahren [85]. Alle Patienten, unabhängig

von ihrem sozioökonomischen Hintergrund, ihrer Ethnie oder ihrem Wohnort, sollten von den Fortschritten in der medizinischen Technologie profitieren können. Nur durch die Schaffung fairer und gerechter Zugangsbedingungen können die Potenziale der KI voll ausgeschöpft und das Gesundheitssysteme verbessert werden.

Um die Weiterentwicklung und Integration von KI in die Medizin zu unterstützen, sind also sorgfältige Planung und umfassende Sicherheitsmaßnahmen unerlässlich. Durch die Einhaltung strenger Leitlinien und die Schaffung transparenter Kommunikationswege kann nicht nur die Effizienz medizinischer Verfahren verbessert werden, sondern auch das Vertrauen der Patienten in diese neuen Technologien gestärkt werden.

5 Zusammenfassung

Die vorliegende Arbeit untersucht die Möglichkeit der Integration von Künstlicher Intelligenz, insbesondere von dem Large Language Model ChatGPT, in den Prozess der medizinischen Dokumentation und der Patientenkommunikation.

Seit der Einführung des Patientenrechtegesetzes im Jahr 2013 und der damit verbundenen Dokumentationspflicht hat der Zeitaufwand für Behandler erheblich zugenommen. Dieser bürokratische Mehraufwand beeinträchtigt nicht nur die Effizienz des medizinischen Arbeitsalltags, sondern auch die Zufriedenheit der Behandler. Gleichzeitig verringert er die Zeit, die für den direkten persönlichen Kontakt mit den Patienten und somit auch zur Entwicklung eines Vertrauensverhältnisses zur Verfügung steht.

Der Einsatz von KI bietet daher ein großes Potenzial, diesen Prozess zu optimieren, indem sprachbasierte Systeme wie ChatGPT repetitive Aufgaben wie die Erstellung von Arztbriefen, Röntgenbefunden oder individuell angepassten Patienteninformationsschriften übernehmen.

Das Ziel dieser Arbeit war es daher, neben der praktischen Anwendbarkeit, der Effizienz und der Qualität der durch ChatGPT generierten Texte auch die Patientenperspektive auf diese neuen Technologien zu beleuchten. Dies erfolgte vor dem Hintergrund der Möglichkeit, den Dokumentationsaufwand der Behandler im klinischen Alltag zu reduzieren und somit den zwischenmenschlichen Kontakt zu verbessern und folglich das Patientenvertrauen zu stärken.

Im Rahmen dieser experimentellen Studie wurden von Zahnmedizinstudierenden formulierte Röntgenberichte mit durch ChatGPT-4 generierten Berichten verglichen. Während die Studierenden die Röntgenbilder zunächst frei und anschließend mithilfe einer standardisierten Checkboxliste beurteilten, erstellte die KI auf Basis dieser ausgefüllten Listen, strukturierte Befundberichte. Die Untersuchung umfasste die Bewertung der Vollständigkeit, sprachlichen Genauigkeit und Lesbarkeit der Texte sowie deren Informationsgehalt. Darüber hinaus wurde untersucht, wie Patienten die Verständlichkeit der KI-generierten Berichte wahrnehmen und bewerten.

Die Ergebnisse zeigten, dass die Nutzung der KI die Bearbeitungszeit signifikant reduzierte, da die Studierenden für die manuelle Formulierung eines Berichts durchschnittlich 30 Minuten benötigten, während das Ausfüllen der Checkboxliste nur 10 Minuten erforderte. Die Checkboxlisten erwiesen sich als hilfreich, um die Vollständigkeit der Dokumentation zu gewährleisten und Flüchtigkeitsfehler zu minimieren. Gleichzeitig wurde deutlich, dass die Qualität der KI generierten Texte stark vom Format der Eingabedaten und der präzisen Gestaltung der Prompts abhängt.

Die Analyse der Berichte ergab, dass die durch ChatGPT generierten Texte frei von Rechtschreib-, Grammatik- und Zeichensetzungsfehlern waren, was die Lesbarkeit und

Verständlichkeit erheblich steigerte. Gleichzeitig war die sprachliche Qualität der KI-Texte vergleichbar mit den Berichten der Studierenden. Allerdings ließ die KI gelegentlich relevante Informationen aus, insbesondere bei komplexeren Datenformaten.

Um die Patientenperspektive einzubeziehen, wurden 150 Patienten gebeten, KI-generierte Röntgenberichte zu bewerten, die auf unterschiedlich formulierten Prompts basierten. Die Akzeptanz variierte dabei deutlich je nach Sprachstil: Patientengerecht formulierte Berichte wurden signifikant positiver bewertet als wissenschaftlich gehaltene Texte. Besonders hervorgehoben wurden die verständliche Erklärung von Fachbegriffen und der respektvolle Ton, was zu besserer Kommunikation und größerem Vertrauen führte. Die Ergebnisse unterstreichen das Potenzial KI-gestützter Anwendungen, nicht nur die Effizienz und Qualität der klinischen Dokumentation zu verbessern, sondern auch die Patientenkommunikation und -aufklärung durch verständlichere und klarer formulierte Texte nachhaltig zu fördern.

Gleichzeitig wurden Herausforderungen deutlich – insbesondere im Hinblick auf die Vollständigkeit der Informationen und die Integration komplexer medizinischer Daten. Zudem stellen ethische und rechtliche Aspekte wie Datenschutz, Diskriminierungsfreiheit und die Klärung der Verantwortlichkeit im Fehlerfall zentrale Voraussetzungen dar, um Vertrauen in die Technologie und deren sichere Anwendung zu gewährleisten.

Der Einsatz von KI im Gesundheitswesen erfordert daher ein engmaschiges Kontrollsystem sowie ein robustes Fehlermanagement. Darüber hinaus sollte die medizinische Ausbildung gezielt angepasst werden, um zukünftige Behandler im kompetenten Umgang mit KI-Systemen zu schulen und ein fundiertes Verständnis für deren Funktionsweise und Grenzen zu vermitteln.

Eine verantwortungsvoll implementierte Integration von Künstlicher Intelligenz in die medizinische Praxis hat das Potenzial, nicht nur die Arbeitsbelastung im klinischen Alltag deutlich zu reduzieren, sondern auch die Qualität der Patientenversorgung und die Kommunikation zwischen Arzt und Patient nachhaltig zu verbessern.

6 Literaturverzeichnis

1. Bayer, T., *Ärztliche Dokumentationspflicht und Einsichtsrecht in Patientenakten: Eine Untersuchung zu den §§ 630f und 630g BGB mit Bezügen zum nationalen sowie europäischen Datenschutzrecht*. 2018: Springer-Verlag.
2. Europe, H., *Auf den Spuren der Zeitdiebe im Krankenhaus: Die wahre Belastung durch Dokumentation an deutschen Akutkrankenhäusern wird unterschätzt*. Im Internet: <http://www.engage.nuance.de/himss-klinische-dokumentation>, 2015.
3. Sönnichsen, A.C., N. Donner-Banzhoff, and E. Baum, *Motive, Berufsziele und Hoffnungen von Studienanfängern im Fach Medizin*. ZFA-Zeitschrift für Allgemeinmedizin, 2005. **81**(05): p. 222-225.
4. Laschet, H., *Ärztmangel bereitet weiter Sorgen*. Schmerzmedizin, 2019. **35**(3): p. 74-74.
5. Jones, O.T., et al., *Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review*. Lancet Digit Health, 2022. **4**(6): p. e466-e476.
6. Attia, Z.I., et al., *Application of artificial intelligence to the electrocardiogram*. Eur Heart J, 2021. **42**(46): p. 4717-4730.
7. Baxi, V., et al., *Digital pathology and artificial intelligence in translational medicine and clinical practice*. Mod Pathol, 2022. **35**(1): p. 23-32.
8. Gore, J.C., *Artificial intelligence in medical imaging*. Magn Reson Imaging, 2020. **68**: p. A1-a4.
9. AI, H., *High-level expert group on artificial intelligence*. 2019, European Commission. Available at: <https://ec.europa.eu/digital-single> p. 6.
10. Arrieta, A.B., et al., *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*. Information fusion, 2020. **58**: p. 82-115.
11. Tretter, M., D. Samhammer, and P. Dabrock, *Künstliche Intelligenz in der Medizin: Von Entlastungen und neuen Anforderungen im ärztlichen Handeln*. Ethik in der Medizin, 2023: p. 1-23.

-
12. Beck, S., M. Faber, and S. Gerndt, *Rechtliche Aspekte des Einsatzes von KI und Robotik in Medizin und Pflege*. Ethik in der Medizin, 2023. **35**(2): p. 247-263.
 13. Zech, H. and I.C. Hünefeld, *Einsatz von KI in der Medizin: Haftung und Versicherung*. Medizinrecht, 2023. **41**(1): p. 1-8.
 14. Ali, S.R., et al., *Using ChatGPT to write patient clinic letters*. Lancet Digit Health, 2023. **5**(4): p. e179-e181.
 15. Wojcik, S., et al., *Beyond ChatGPT: What does GPT-4 add to healthcare? The dawn of a new era*. Cardiol J, 2023. **30**(6): p. 1018-1025.
 16. Kalyan, K.S., *A survey of GPT-3 family large language models including ChatGPT and GPT-4*. Natural Language Processing Journal, 2024. **6**: p. 100048.
 17. Zong, M. and B. Krishnamachari, *A survey on GPT-3*. arXiv preprint arXiv:2212.00857, 2022.
 18. Taecharungroj, V., *“What can ChatGPT do?” Analyzing early reactions to the innovative AI chatbot on Twitter*. Big Data and Cognitive Computing, 2023. **7**(1): p. 35.
 19. Cascella, M., et al., *Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios*. J Med Syst, 2023. **47**(1): p. 33.
 20. Albrecht, S., *ChatGPT und andere Computermodelle zur Sprachverarbeitung–Grundlagen, Anwendungspotenziale und mögliche Auswirkungen*. 2023.
 21. Mahesh, B., *Machine learning algorithms-a review*. International Journal of Science and Research (IJSR).[Internet], 2020. **9**(1): p. 381-386.
 22. Wörrlein, B., et al., *Deep-Learning-basierte Prognose von Stromverbrauch für die hybride Simulation*. Simulation in Produktion und Logistik 2019, 2019: p. 121-131.
 23. Vaswani, A., et al., *Attention is all you need*. Advances in neural information processing systems, 2017. **30**.
 24. Hashana, A.J., et al. *Deep Learning in ChatGPT-A Survey*. in *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*. 2023. IEEE.
 25. De Angelis, L., et al., *ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health*. Front Public Health, 2023. **11**: p. 1166120.

-
26. Ghojogh, B. and A. Ghodsi, *Attention mechanism, transformers, BERT, and GPT: tutorial and survey*. 2020.
 27. Cho, K., et al., *On the properties of neural machine translation: Encoder-decoder approaches*. arXiv preprint arXiv:1409.1259, 2014.
 28. Krüger, R., *Die Transformer-Architektur für Systeme zur neuronalen maschinellen Übersetzung—eine popularisierende Darstellung*. 2021, trans-kom.
 29. Staudemeyer, R.C. and E.R. Morris, *Understanding LSTM—a tutorial into long short-term memory recurrent neural networks*. arXiv preprint arXiv:1909.09586, 2019.
 30. Hu, Y., et al., *Overcoming the vanishing gradient problem in plain recurrent networks*. arXiv preprint arXiv:1801.06105, 2018.
 31. Le, P. and W. Zuidema, *Quantifying the vanishing gradient and long distance dependency problem in recursive neural networks and recursive LSTMs*. arXiv preprint arXiv:1603.00423, 2016.
 32. Hochreiter, S. and J. Schmidhuber, *LSTM can solve hard long time lag problems*. Advances in neural information processing systems, 1996. **9**.
 33. Pham, N.-Q., et al., *Very deep self-attention networks for end-to-end speech recognition*. arXiv preprint arXiv:1904.13377, 2019.
 34. Brown, T., et al., *Language models are few-shot learners*. Advances in neural information processing systems, 2020. **33**: p. 1877-1901.
 35. Luccioni, A.S. and J.D. Viviano, *What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus*. arXiv preprint arXiv:2105.02732, 2021.
 36. Moor, M., et al., *Foundation models for generalist medical artificial intelligence*. Nature, 2023. **616**(7956): p. 259-265.
 37. Kirk, H.R., et al., *Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models*. Advances in neural information processing systems, 2021. **34**: p. 2611-2624.
 38. Fang, X., et al., *Bias of AI-generated content: an examination of news produced by large language models*. Scientific Reports, 2024. **14**(1): p. 1-20.

-
39. Logé, C., et al., *Q-Pain: a question answering dataset to measure social bias in pain management*. arXiv preprint arXiv:2108.01764, 2021.
 40. Hoffman, K.M., et al., *Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites*. Proceedings of the National Academy of Sciences, 2016. **113**(16): p. 4296-4301.
 41. Shaier, S., et al., *Emerging Challenges in Personalized Medicine: Assessing Demographic Effects on Biomedical Question Answering Systems*. arXiv preprint arXiv:2310.10571, 2023.
 42. Memon, S.A. and J.D. West, *Search engines post-ChatGPT: How generative artificial intelligence could make search less reliable*. arXiv preprint arXiv:2402.11707, 2024.
 43. Yao, J.-Y., et al., *Llm lies: Hallucinations are not bugs, but features as adversarial examples*. arXiv preprint arXiv:2310.01469, 2023.
 44. White, J., et al., *Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design*. arXiv preprint arXiv:2303.07839, 2023.
 45. Heston, T.F. and C. Khun, *Prompt engineering in medical education*. International Medical Education, 2023. **2**(3): p. 198-205.
 46. Zhou, Y., et al., *Large language models are human-level prompt engineers*. arXiv preprint arXiv:2211.01910, 2022.
 47. Kung, T.H., et al., *Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models*. PLoS digital health, 2023. **2**(2): p. e0000198.
 48. Takagi, S., et al., *Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study*. JMIR Medical Education, 2023. **9**(1): p. e48002.
 49. Achiam, J., et al., *Gpt-4 technical report*. arXiv preprint arXiv:2303.08774, 2023.
 50. Nakano, R., et al., *Webgpt: Browser-assisted question-answering with human feedback*. arXiv preprint arXiv:2112.09332, 2021.
 51. Low, A. and Z.Y. Kalender, *Data dialogue with chatgpt: Using code interpreter to simulate and analyse experimental data*. arXiv preprint arXiv:2311.12415, 2023.

-
52. Patel, S.B. and K. Lam, *ChatGPT: the future of discharge summaries?* The Lancet Digital Health, 2023. **5**(3): p. e107-e108.
 53. Waisberg, E., et al., *GPT-4 and Ophthalmology Operative Notes*. Ann Biomed Eng, 2023. **51**(11): p. 2353-2355.
 54. Liao, W., et al., *Differentiating ChatGPT-Generated and Human-Written Medical Texts: Quantitative Study*. JMIR Med Educ, 2023. **9**: p. e48904.
 55. Baker, H.P., et al., *ChatGPT's Ability to Assist with Clinical Documentation: A Randomized Controlled Trial*. JAAOS-Journal of the American Academy of Orthopaedic Surgeons, 2024. **32**(3): p. 123-129.
 56. Bär, T., *Die spontane Gesprächszeit von Patienten zu Beginn des Arztgesprächs in der hausärztlichen Praxis*. 2009.
 57. Theisel, S., T. Schielein, and H. Spießl, *Der „ideale“ Arzt aus Sicht psychiatrischer Patienten*. Psychiatrische Praxis, 2010. **37**(06): p. 279-284.
 58. Bachmaier, S., *Grundpfeiler für eine gelingende Arzt-Patienten-Kommunikation*. Im Fokus Onkologie, 2021. **24**(Suppl 1): p. 62-64.
 59. Muscat, D.M., et al., *Health literacy and shared decision-making: exploring the relationship to enable meaningful patient engagement in healthcare*. Journal of general internal medicine, 2021. **36**: p. 521-524.
 60. Rooney, M.K., et al., *Readability of patient education materials from high-impact medical journals: a 20-year analysis*. Journal of patient experience, 2021. **8**: p. 2374373521998847.
 61. Liu, C., et al., *What is the meaning of health literacy? A systematic review and qualitative synthesis*. Family medicine and community health, 2020. **8**(2).
 62. Wang, J., et al., *Prompt engineering for healthcare: Methodologies and applications*. arXiv preprint arXiv:2304.14670, 2023.
 63. Graafen, M.A., M. Sennekamp, and A. Messemaker, *„Also, im Internet steht...“—Wenn Ärztinnen und Ärzte auf internetinformierte Patienten treffen*. Zeitschrift für Allgemeinmedizin, 2021. **97**(5): p. 210-214.

-
64. Spiegel, J.L., et al., *Beurteilung von im Internet verfügbaren deutschsprachigen Patienteninformationen zum Hörsturz*. *Laryngo-Rhino-Otologie*, 2021. **100**(08): p. 618-626.
 65. Paul, S., et al., *Systematic Analysis of Readability of Patient Information on Internet Pages from Departments for Trauma Surgery of German University Hospitals*. *Zeitschrift für Orthopädie und Unfallchirurgie*, 2021. **159**(02): p. 187-192.
 66. Hendy, A., et al., *How good are gpt models at machine translation? a comprehensive evaluation*. arXiv preprint arXiv:2302.09210, 2023.
 67. (Destatis), S.B. *Ausländische Bevölkerung nach ausgewählten Staatsangehörigkeiten 2024* [cited 2024 09.05.2024]; Available from: <https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Migration-Integration/Tabellen/auslaendische-bevoelkerung-staatsangehoerigkeit-jahre.html>.
 68. Juravle, G., et al., *Trust in artificial intelligence for medical diagnoses*. *Progress in brain research*, 2020. **253**: p. 263-282.
 69. Ferrario, A. and M. Loi. *How explainability contributes to trust in AI*. in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022.
 70. Alam, L. and S. Mueller, *Examining the effect of explanation on satisfaction and trust in AI diagnostic systems*. *BMC medical informatics and decision making*, 2021. **21**(1): p. 178.
 71. Von Eschenbach, W.J., *Transparency and the black box problem: Why we do not trust AI*. *Philosophy & Technology*, 2021. **34**(4): p. 1607-1622.
 72. Dave, T., S.A. Athaluri, and S. Singh, *ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations*. *Front Artif Intell*, 2023. **6**: p. 1169595.
 73. Sonntagbauer, M., M. Haar, and S. Kluge, *[Artificial intelligence: How will ChatGPT and other AI applications change our everyday medical practice?]*. *Med Klin Intensivmed Notfmed*, 2023. **118**(5): p. 366-371.
 74. Flesch, R., *A new readability yardstick*. *Journal of Applied Psychology*, 1948. **32**(3): p. 221-233.
 75. Amstad, T., *Wie verständlich sind unsere Zeitungen?* 1978: Studenten-Schreib-Service.

-
76. Björnsson, C.-H., *Readability of newspapers in 11 languages*. Reading Research Quarterly, 1983: p. 480-497
 77. Nickel, S., *Textschwierigkeit objektivieren: Der Lesbarkeitsindex LIX – Wie schwierig sind Lesetexte in der Alphabetisierung?* 2011: p. 30-32.
 78. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805, 2018.
 79. Deutsch, T., M. Jasbi, and S. Shieber, *Linguistic features for readability assessment*. arXiv preprint arXiv:2006.00377, 2020.
 80. Bellinger, J.R., et al., *BPPV information on google versus AI (ChatGPT)*. Otolaryngology–Head and Neck Surgery, 2023.
 81. Deveci, C.D., et al., *A comparison of cover letters written by ChatGPT-4 or humans*. Dan Med J, 2023. **70**(12): p. A06230412.
 82. Rajpurohit, S., et al., *Development and evaluation of patient information leaflet for liver cirrhosis patients*. Clinical Epidemiology and Global Health, 2023. **24**: p. 101436.
 83. Storesund, A., et al., *Clinical efficacy of combined surgical patient safety system and the world health organization's checklists in surgery: a nonrandomized clinical trial*. Jama surgery, 2020. **155**(7): p. 562-570.
 84. Weichert, T., *Datenschutz im Kontext der medizinischen Nutzung von KI-Systemen: heute und morgen*. Zeitschrift für medizinische Ethik, 2021. **67**(3): p. 351-365.
 85. Union, E.U.C.d.G.d.E., *Charta der Grundrechte der Europäischen Union*, in *TITEL III- Gleichheit*. 14.12.2007.
 86. Haftenberger, A. and C. Dierks, *Rechtliche Einordnung von künstlicher Intelligenz in der Inneren Medizin*. Die Innere Medizin, 2023. **64**(11): p. 1044-1050.
 87. Allen, B., et al., *Democratizing ai*. Journal of the American College of Radiology, 2019. **16**(7): p. 961-963.
 88. Zhao, G., et al., *The comparison of teaching efficiency between virtual reality and traditional education in medical education: a systematic review and meta-analysis*. Annals of translational medicine, 2021. **9**(3).

7 Abbildungsverzeichnis

- ABBILDUNG 1: ORTHOPANTOMOGRAMM A. NEBEN EINEM DEUTLICH REDUZIERTEM UND TEILWEISE PROTHETISCH REHABILITIERTEN RESTZAHNBESTAND KANN MAN ZUM EINEN EIN ARTEFAKT EINES METALLISCHEN PIERCINGS ERKENNEN. ZUM ANDEREN LÄSST SICH IM BEREICH DES LINKEN RAMUS MANDIBULAE EIN TRANSLUZENT, OSTEOLYTISCH ERSCHEINENDES AREAL BESCHREIBEN. SOWOHL IM OBER- ALS AUCH IM UNTERKIEFER KANN EIN GENERALISIERTER HORIZONTALER KNOCHENABBAU ERKANNT WERDEN. 15
- ABBILDUNG 2: ORHTOPANTOMOGRAMM B. NEBEN DEN KONSERVIERENDEN BEFUNDEN WIE FÜLLUNGEN UND WURZELKANALBEHANDLUNGEN LASSEN SICH AUCH CHIRURGISCH-PROTHETISCHE BEFUNDE WIE IMPLANTATE MIT SIGNIFIKANTEM PERIIMPLANTÄREN KNOCHENABBAU UND EINE WURZELSPITZENRESEKTION AM ZAHN 46 BESCHREIBEN. AUCH HIER LÄSST SICH EIN METALLISCHES ARTEFAKT EINER RETROGRADEN WURZELFÜLLUNG AM ZAHN 13 ERKENNEN. 15
- ABBILDUNG 3: DIE DARSTELLUNG ZEIGT EIN BEISPIEL EINER DURCH EINEN STUDIERENDEN AUSGEFÜLLTE CHECKBOXLISTE. ES WURDEN DREI (DURCH DIE NUMMERN 1-3 MARKIERTE) TABELLENBLÄTTER ZUR BEFUNDUNG DER OPGs ZUR VERFÜGUNG GESTELLT. DIE KREUZE MARKIEREN DIE INFORMATIONEN, DIE DER STUDIERENDE ALS POSITIV (IM OPG VORHANDEN) MARKIERT HAT UND IM ANSCHLUSS IN DEN DURCH CHATGPT GENERIERTEN BEFUND AUFGENOMMEN WERDEN SOLLEN. 16
- ABBILDUNG 4: FRAGEBOGEN MIT 11 FRAGEN ZUR EVALUATION VON VERSTÄNDLICHKEIT, STRUKTUR UND INFORMATIONSGEHALT, EMPATHIE UND TON SOWIE MOTIVATION UND ZUKÜNFTIGER PERSPEKTIVE DER ORIGINALEN KI-GENERIERTEN UND VON CHATGPT VEREINFACHTEN RÖNTGENBERICHTE. DIE BEWERTUNG ERFOLGTE AUF EINER FÜNFSTUFIGEN LIKERT-SKALA VON 1 („TRIFFT VOLL ZU“) BIS 5 („TRIFFT GAR NICHT ZU“) MIT DEN DAZWISCHEN LIEGENDEN MÖGLICHKEITEN „TRIFFT EHER ZU“ (2), „WEDER NOCH“ (3) UND „TRIFFT EHER NICHT ZU“ (4). 23
- ABBILDUNG 5: LESBARKEIT VON KI-GENERIERTEN RÖNTGENBERICHTEN IM VERGLEICH ZU STUDENTISCH VERFASSTEN RADIOLOGISCHEN BERICHTEN SEMESTERUNABHÄNGIG (A) UND SEMESTERABHÄNGIG (B), BEWERTET MIT DEM FRE-SCORE. DIE DATEN REPRÄSENTIEREN DEN MITTELWERT +/- STANDARDABWEICHUNG. STICHPROBENGRÖßE: N = 100 (A); SEMESTER 1: N=20, SEMESTER 2: N=19, SEMESTER 3: N=21, SEMESTER 4: N=20, SEMESTER 5: N=20 (B); * = P<0,05. 27
- ABBILDUNG 6: LESBARKEIT VON KI-GENERIERTEN RÖNTGENBERICHTEN IM VERGLEICH ZU VON STUDIERENDEN VERFASSTEN RADIOLOGISCHEN BERICHTEN SEMESTERUNABHÄNGIG (A) UND SEMESTERABHÄNGIG (B), BEWERTET MIT DEM LIX-SCORE. DIE DATEN REPRÄSENTIEREN DEN MITTELWERT +/- STANDARDABWEICHUNG. STICHPROBENGRÖßE: N = 100 (A); SEMESTER 1: N=20, SEMESTER 2: N=19, SEMESTER 3: N=21, SEMESTER 4: N=20, SEMESTER 5: N=20 (B); * = P<0,05. 28
- ABBILDUNG 7: BEWERTUNG DER ÄHNLICHKEIT DER KI-GENERIERTEN RÖNTGENBERICHTE IM VERGLEICH ZU REFERENZTEXTEN UNTER VERWENDUNG DES BERT-SCORES (A) MIT PRECISION (P), RECALL (R) UND F1-SCORE (F1). DIE GENAUIGKEIT DER KI-GENERIERTEN RADIOLOGISCHEN BERICHTE WURDE WEITERHIN ÜBER DIE ENTHALTENEN DIAGNOSTISCHEN INFORMATIONEN FÜR DIE GESAMTE CHECKBOXLISTE (B) SOWIE INDIVIDUELL FÜR JEDES ARBEITSBLATT DER CHECKBOXLISTE BEWERTET. DIE DATEN REPRÄSENTIEREN DEN MITTELWERT +/- STANDARDABWEICHUNG. STICHPROBENGRÖßE: N = 100. * = P<0,05. 29
- ABBILDUNG 8: ANALYSE DER WORT- (A), SATZ- (B), SILBEN- (C), DIPHTHONG- (D) UND ZEICHENANZAHL (E) VON KI-GENERIERTEN IM VERGLEICH ZU VON STUDIERENDEN

VERFASSTEN RÖNTGENBERICHTEN. DIE DATEN REPRÄSENTIEREN DEN MITTELWERT +/- STANDARDABWEICHUNG. STICHPROBENGRÖßE: N = 100. * = P<0,05.....	30
ABBILDUNG 9: ANALYSE DER SATZLÄNGE UND DES ANTEILS LANGER WÖRTER (MEHR ALS SECHS BUCHSTABEN) IN KI-GENERIERTEN RÖNTGENBERICHTEN IM VERGLEICH ZU VON STUDIERENDEN VERFASSTEN BERICHTEN. DIE DATEN REPRÄSENTIEREN DEN MITTELWERT +/- STANDARDABWEICHUNG. STICHPROBENGRÖßE: N = 100. *P<0,05 VS. STUDIERENDE.	30
ABBILDUNG 10: ANALYSE DER FEHLERANZAHL (A), EINSCHLIEßLICH GRAMMATIK, RECHTSCHREIBUNG UND ZEICHENSETZUNG, SOWIE DES FEHLERQUOTIENTEN (B), BERECHNET ALS FEHLER / WÖRTER * 100, VON KI-GENERIERTEN RÖNTGENBERICHTEN IM VERGLEICH ZU VON STUDIERENDEN VERFASSTEN BERICHTEN. DIE DATEN REPRÄSENTIEREN DEN MITTELWERT +/- STANDARDABWEICHUNG. STICHPROBENGRÖßE: N = 100. * = P<0,05.....	31
ABBILDUNG 11: ANALYSE DER FRE-WERTE DER VON KI-GENERIERTEN UND VON KI-VEREINFACHTEN RÖNTGENBERICHTE IM VERGLEICH ZU VON STUDIERENDEN VERFASSTEN BERICHTEN INSGESAMT (A) UND EINZELN FÜR JEDES SEMESTER (B), BEWERTET MIT DEM FRE-SCORE. DABEI WURDEN DIE KI-GENERIERTE BERICHTE MIT ZWEI WEITEREN PROMPTS (PROMPT 2 UND PROMPT 3) IN VEREINFACHTEN VERSIONEN UMFORMULIERT, UM DIE LESBARKEIT ZU VERBESSERN. DIE DATEN REPRÄSENTIEREN DEN MITTELWERT ± STANDARDABWEICHUNG. STICHPROBENGRÖßE: N = 100; SEMESTER 1: N=20, SEMESTER 2: N=19, SEMESTER 3: N=21, SEMESTER 4: N=20, SEMESTER 5: N=20. * = P<0,05.....	32
ABBILDUNG 12: ANALYSE DER DURCHSCHNITTLICHEN SATZANZAHL (A), WORTANZAHL (B) UND DES ANTEILS LANGER WÖRTER (C) DER VON KI-GENERIERTEN UND VON KI-VEREINFACHTEN RÖNTGENBERICHTE IM VERGLEICH ZU VON STUDIERENDEN VERFASSTEN BERICHTEN. DABEI WURDEN DIE KI-GENERIERTE BERICHTE MIT ZWEI WEITEREN PROMPTS (PROMPT 2 UND PROMPT 3) IN VEREINFACHTEN VERSIONEN UMFORMULIERT, UM DIE LESBARKEIT ZU VERBESSERN. DIE DATEN REPRÄSENTIEREN DEN MITTELWERT ± STANDARDABWEICHUNG. STICHPROBENGRÖßE: N = 100. * = P<0,05.....	33
ABBILDUNG 13: BEWERTUNGEN DES PATIENTENVERSTÄNDNISSES DER RÖNTGENBEFUNDE FÜR DIE ORIGINALEN KI-GENERIERTEN RÖNTGENBERICHTE (TEXT 1 – ROT), DIE VEREINFACHTEN VERSION (TEXT 2 – BLAU) SOWIE DIE STARK VEREINFACHTEN VERSION (TEXT 3 – GRÜN). EVALUIERT WURDEN DIE ERKLÄRUNGEN DER VERWENDETEN FACHBEGRIFFE (A) SOWIE DIE VERSTÄNDLICHKEIT OHNE ZUSÄTZLICHE HILFE (B). STICHPROBENGRÖßE: N=50; * = P<0,05.	35
ABBILDUNG 14: BEWERTUNGEN DES PATIENTENVERSTÄNDNISSES DER RÖNTGENBEFUNDE FÜR DIE ORIGINALEN KI-GENERIERTEN RÖNTGENBERICHTE (TEXT 1 – ROT), DIE VEREINFACHTEN VERSION (TEXT 2 – BLAU) SOWIE DIE STARK VEREINFACHTEN VERSION (TEXT 3 – GRÜN). EVALUIERT WURDEN DIE STRUKTUR DES RÖNTGENBERICHTS (A) SOWIE DIE DETAILLIERTHEIT DER VERMITTELTEN INFORMATIONEN (B). STICHPROBENGRÖßE: N=50; * = P<0,05.	35
ABBILDUNG 15: BEWERTUNGEN DES PATIENTENVERSTÄNDNISSES DER RÖNTGENBEFUNDE FÜR DIE ORIGINALEN KI-GENERIERTEN RÖNTGENBERICHTE (TEXT 1 – ROT), DIE VEREINFACHTEN VERSION (TEXT 2 – BLAU) SOWIE DIE STARK VEREINFACHTEN VERSION (TEXT 3 – GRÜN). TON DES RÖNTGENBERICHTS (A), SOWIE DIE EMPATHISCHE HALTUNG GEGENÜBER DEM VERSTÄNDNISNIVEAU (B). STICHPROBENGRÖßE: N=50; * = P<0,05.....	36
ABBILDUNG 16: BEWERTUNGEN DES PATIENTENVERSTÄNDNISSES DER RÖNTGENBEFUNDE FÜR DIE ORIGINALEN KI-GENERIERTEN RÖNTGENBERICHTE (TEXT 1 – ROT), DIE VEREINFACHTEN VERSION (TEXT 2 – BLAU), SOWIE DIE STARK VEREINFACHTEN VERSION (TEXT 3 – GRÜN). EVALUIERT WURDE, OB DER RÖNTGENBERICHT HILFT IN WEITEREN GESPRÄCHEN DIE RICHTIGEN FRAGEN ZU STELLEN (A), SOWIE OB DIE MÖGLICHKEIT EIN WEITERES GESPRÄCH AUF AUGENHÖHE ZU FÜHREN BESTEHT (B). STICHPROBENGRÖßE: N=50; * = P<0,05.	37

-
- ABBILDUNG 17: BEWERTUNGEN DES PATIENTENVERSTÄNDNISSES DER RÖNTGENBEFUNDE FÜR DIE ORIGINALEN KI-GENERIERTEN RÖNTGENBERICHTE (TEXT 1 – ROT), DIE VEREINFACHTE VERSION (TEXT 2 – BLAU) SOWIE DIE STARK VEREINFACHTE VERSION (TEXT 3 – GRÜN). EVALUIERT WURDEN, OB DER RÖNTGENBERICHT LEICHT ZU VERSTEHEN SEI (A) SOWIE DIE MOTIVATION SICH ZUKÜNFTIG MEHR MIT MUNDHYGIENE UND GESUNDHEIT AUSEINANDERZUSETZEN (B). STICHPROBENGROÖßE: N=50; * = P<0,05. 37
- ABBILDUNG 18: BEWERTUNGEN DES PATIENTENVERSTÄNDNISSES DER RÖNTGENBEFUNDE FÜR DIE ORIGINALEN KI-GENERIERTEN RÖNTGENBERICHTE (TEXT 1 – ROT), DIE VEREINFACHTE VERSION (TEXT 2 – BLAU) SOWIE DIE STARK VEREINFACHTE VERSION (TEXT 3 – GRÜN). EVALUIERT WURDE, OB ALLE MEDIZINISCHEN BERICHTE AUF DIESEM VERSTÄNDNISNIVEAU GESCHRIEBEN SEIN SOLLTEN. STICHPROBENGROÖßE: N=50; * = P<0,05. 38
- ABBILDUNG 19: DIE ABBILDUNG ZEIGT DAS INTERNATIONALE ZAHNSCHEMA DER FDI. DIE ZÄHNE SIND HIER IN VIER QUADRANTEN AUFGETEILT. DIE ERSTE ZAHL BESCHREIBT DEN QUADRANTEN, DIE ZWEITE ZAHL BESCHREIBT DIE POSITION DES ZAHNES. QUELLE: [HTTPS://WWW.ZAHNE.NET/DIE-BEZEICHNUNG-DER-ZAHNE](https://www.zahne.net/die-bezeichnung-der-zahne) (25.05.2025) 68
- ABBILDUNG 20: DIE ABBILDUNG ZEIGT EINEN AUSSCHNITT AUS DER EXCEL-DATEI MIT DEM DURCH EINEN STUDIERENDEN EXEMPLARISCH AUSGEFÜLLTEN TABELLENBLATT, WELCHES DEN TITEL DER WORD-DATEI „PROTHETISCHE BEFUNDE“ ERHALTEN HAT. DIE CHECKBOX-SPALTE WURDE ALS „CHECKBOX“ BETITELT UND SO IM PROMPT BESCHRIEBEN. DIE IM RÖNTGENBERICHT ZU NENNENDEN AUSSAGEN WURDEN MIT „X“ MARKIERT. 69
- ABBILDUNG 21: DIE ABBILDUNG ZEIGT DIE ZWEI TABELLENBLÄTTER „SAPIENTES“ (A) UND „ZYSTISCHE VERÄNDERUNGEN“ (B), WELCHE SICH DEUTLICH IM AUFBAU UNTERSCHIEDEN, WAS ZU FEHLINTERPRETATIONEN DER KI GEFÜHRT HABEN KÖNNTE. ES WIRD GEZEIGT, WIE DIE EINZELNEN TABELLENBLÄTTER EXEMPLARISCH AUSGEFÜLLT WORDEN SEIN KÖNNTEN. DAS „X“ MARKIERT HIERBEI EINEN POSITIVEN BEFUND, DER IN DEN RÖNTGENBERICHT AUFGENOMMEN WERDEN SOLLTE. 70

8 Tabellenverzeichnis

TABELLE 1: VERGLEICHSWERTE DES FLESCHEADING EASE-SCORE	19
TABELLE 2: VERGLEICHSWERTE DES LESBARKEITSINDEX (LIX)	20

9 Anhang

9.1 Röntgenberichte der Patientenevaluation

Im Folgenden sind die drei randomisiert ausgewählten KI-generierten Texte für die drei unterschiedlichen Prompts abgebildet, die den Patienten zur Bewertung und Evaluation vorgelegt worden sind.

9.1.1 Prompt 1:

„Die Positionierung des Patienten im OPG ist regelrecht und die Belichtung ist regelhaft. Die Ramusstrukturen beiderseits erscheinen unauffällig. Es zeigt sich eine unilaterale Verschattung der rechten Kieferhöhle, während die Kondylen des Kiefergelenks unauffällig sind. Der allgemeine Zahnstatus ist sowohl konservierend als auch prothetisch gut versorgt. Artefakte sind durch ein Piercing im Bereich der Nase und Lippe zu erkennen. Prothetische Befunde weisen auf Brücken im ersten und zweiten Quadranten hin. Insbesondere die Zähne 16, 14, 25 und 27 sind prothetisch versorgt, die Versorgungen scheinen ohne besonderen Befund zu sein. Eine Wurzelkanalbehandlung ist an Zahn 14 durchgeführt worden. Die fehlenden Zähne sind 17, 15, 26, 28, 38, 37, 36, 35, 46, 47 und 48. Zudem zeigen sich apikale Aufhellungen an den Zähnen 16 und 27. Es sind keine Implantate vorhanden. Der Knochen im Ober- und Unterkiefer zeigt einen generellen horizontalen Knochenverlust. Es sind zystische Veränderungen im linken Ramus vorhanden, die sowohl scharf als auch unscharf begrenzt sind, jedoch ohne Kontinuitätsunterbrechung der Compacta. Aufgrund der Befunde empfehle ich eine weitere 3D-Bildgebung, etwa durch DVT, MRT oder CT.“

9.1.2 Prompt 2:

„Liebe Patientin, lieber Patient,

bei der kürzlich durchgeführten Röntgenaufnahme Ihres Mund- und Kieferbereichs haben wir festgestellt, dass Ihr Kopf leicht zur Seite geneigt war, als die Aufnahme gemacht wurde. Trotzdem ist die Qualität des Bildes gut, und alle wichtigen Strukturen sind klar zu erkennen.

Ihre Kieferknochen, einschließlich der Bereiche, die sich um die Gelenke herum befinden, sowie Ihre Kieferhöhlen, sehen normal aus. Das ist eine gute Nachricht! Wir haben gesehen, dass Ihre Zähne gut versorgt sind, sowohl was Füllungen als auch Zahnprothesen betrifft. Es gibt jedoch ein paar Auffälligkeiten, die durch ein Piercing im Bereich Ihrer Nase oder Lippe verursacht wurden, aber das ist nichts Beunruhigendes.

Bezüglich der Zahnprothesen haben wir Brücken im oberen linken (zweiten Quadranten) und oberen rechten Bereich (ersten Quadranten) Ihres Mundes bemerkt. Einige Zähne, nämlich

die Nummern 16, 14, 25 und 27, sind mit diesen Prothesen nicht optimal versorgt. Der Zahn Nummer 14 wurde nach allen Regeln der Kunst wurzelbehandelt, und der Zahn Nummer 33 hat eine Füllung. Wir haben auch Kronen auf den Zähnen Nummer 16, 15, 14, 25, 26 und 27 festgestellt, was zeigt, dass diese Zähne gut geschützt sind.

Es gibt einige Zähne, die fehlen – speziell die Nummern 18, 28, 38, 37, 36, 35, 46, 47 und 48. Aber das ist in Ihrem Röntgenbild bereits vermerkt. Bei einem Zahn im unteren vorderen Bereich, Nummer 41, haben wir eine dunkle Stelle entdeckt, die eine weitere Untersuchung benötigt, da sie auf einen möglichen Hohlraum unterhalb der Zahnwurzel hinweisen könnte.

Es gibt keine Anzeichen für Zahnimplantate, was für die Beurteilung Ihres Mundgesundheitszustands wichtig ist. Der Knochen im Unterkiefer zeigt insgesamt eine Abnahme der Dichte, was auf einen allgemeinen Knochenabbau hindeutet. Besonders in einigen Bereichen des Oberkiefers haben wir einen verstärkten Knochenabbau festgestellt.

Außerdem haben wir eine Veränderung im Bereich des linken Kieferwinkels bemerkt, die aussieht wie eine Zyste mit mehreren Kammern und einer Unterbrechung der normalen Knochenstruktur. Dies sollte weiter untersucht werden.

Als nächste Schritte empfehlen wir, in 6 Monaten eine weitere Röntgenaufnahme zu machen, um die Entwicklung zu überwachen. Es könnte auch hilfreich sein, eine 3D-Aufnahme Ihres Kiefers machen zu lassen, um einen genaueren Blick auf bestimmte Bereiche zu werfen. Abhängig von den Ergebnissen könnten wir eine kleine Operation in Betracht ziehen, um das Problem mit der Zyste zu lösen, sowie weitere Untersuchungen, um sicherzustellen, dass alles in Ordnung ist.

Wir hoffen, diese Informationen helfen Ihnen, den Bericht besser zu verstehen. Wenn Sie Fragen haben oder weitere Erläuterungen benötigen, stehen wir Ihnen gerne zur Verfügung.

Mit freundlichen Grüßen,

Ihr Zahnarztteam“

9.1.3 Prompt 3:

„Nachdem wir Ihr Röntgenbild, das einen umfassenden Blick auf Ihren Mund- und Kieferbereich bietet, genau angeschaut haben, hier eine verständliche Zusammenfassung der Ergebnisse:

Ihr Kopf war beim Röntgen leicht zur Seite geneigt, aber keine Sorge, das Bild ist trotzdem sehr klar. Die Knochenstruktur in beiden Seiten Ihres Unterkiefers und die Gelenke, die Ihren Kiefer bewegen lassen, sehen normal aus. Allerdings haben wir eine dunklere Schattierung in den Hohlräumen oberhalb Ihrer Zähne festgestellt, die auf beiden Seiten gleich ist.

Was die Gesundheit Ihrer Zähne betrifft: Sie sind gut versorgt, sowohl mit Schutzbehandlungen als auch mit Zahnersatz. Allerdings gibt es einige Zähne, die wir im Auge behalten sollten. Sie haben Zahnersatz in Form von künstlichen Zahnbrücken im hinteren Bereich Ihres Mundes, besonders auf den Zähnen 37 und 36. Bei den Zähnen 17, 13 und 46 wurden spezielle Behandlungen durchgeführt, die das Innere der Zähne betreffen, und diese sehen gut aus. Füllungen finden sich in den Zähnen 12, 11, 21, 22 und 46. Zudem haben die Zähne 17, 16, 13, 26, 37, 36, 35 und 34 künstliche "Kappen", die wir Kronen nennen.

Einige Ihrer Zähne, nämlich 15, 25 und 45, sind nicht mehr vorhanden. Bei Zahn 46 haben wir einen hellen Bereich entdeckt, der unsere Aufmerksamkeit erfordert. An den Stellen, wo die Zähne 37, 36 und 34 fehlen, wurden künstliche Zahnwurzeln, sogenannte Implantate, eingesetzt. Rund um diese Implantate haben wir Knochenverlust festgestellt.

Generell gibt es einen Rückgang des Knochens, der Ihre Zähne im Kiefer hält, sowohl oben als auch unten. Besonders in den Bereichen der Implantate müssen wir den Knochenverlust genau beobachten.

Als nächste Schritte empfehlen wir, weitere Untersuchungen durchzuführen, um sicherzustellen, dass wir alles für Ihre Mundgesundheit tun.

Bitte denken Sie daran, dass diese Zusammenfassung auf Ihrem Röntgenbild basiert. Weitere Untersuchungen sind nötig, um einen vollständigen Überblick zu bekommen und den besten Behandlungsplan für Sie zu erstellen.“

9.2 Europäisches Zahnschema

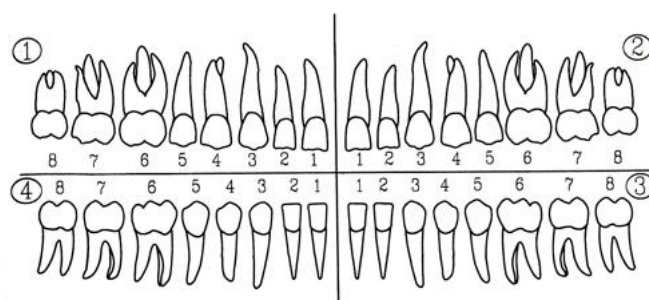


Abbildung 19: Die Abbildung zeigt das internationale Zahnschema der FDI. Die Zähne sind hier in vier Quadranten aufgeteilt. Die erste Zahl beschreibt den Quadranten, die zweite Zahl beschreibt die Position des Zahnes. Quelle: <https://www.zahne.net/die-bezeichnung-der-zahne> (25.05.2025)

9.3 Beispiele der Excel-Tabelle

	A	B	C	D	E	F
1	Checkbox					
2	X	Brücken im ersten Quadranten				
3	X	Brücken im zweiten Quadranten				
4		Brücken im dritten Quadranten				
5		Brücken im vierten Quadranten				
6						
7		Zähne, die Brückenanker darstellen				
8		18				
9		17				
10	X	16				
11		15				
12	X	14				
13		13				
14		12				
15		11				
16		21				
17		22				
18		23				
19		24				
20	X	25				
21		26				
22	X	27				
23		28				
24		38				
25		37				
26		36				
27		35				
28		34				
29		33				
30		32				
31		31				
32		41				
33		42				
34		43				
35		44				
36		45				
37		46				
38		47				
39		48				
40						
41	X	insuffizient				
42		unauffällig				
43		keine Brücken vorhanden				
44						
45						
46						
47						
48						
49						
50						
51						

Abbildung 20: Die Abbildung zeigt einen Ausschnitt aus der Excel-Datei mit dem durch einen Studierenden exemplarisch ausgefüllten Tabellenblatt, welches den Titel der Word-Datei „Prothetische Befunde“ erhalten hat. Die Checkbox-Spalte wurde als „Checkbox“ betitelt und so im Prompt beschrieben. Die im Röntgenbericht zu nennenden Aussagen wurden mit „X“ markiert.

A		A	B	C	D
			Checkbox		
		18	<input checked="" type="checkbox"/>	Durchgebrochen	
3		18		im Durchbruch	
4		18		angelegt	
5		18		Nervkanalnähe	
6		18		verlagert	
7		18		retiniert	
8					
9		28		Durchgebrochen	
10		28		im Durchbruch	
11		28		angelegt	
12		28		Nervkanalnähe	
13		28		verlagert	
14		28		retiniert	
15					
16		38		Durchgebrochen	
17		38		im Durchbruch	
18		38		angelegt	
19		38		Nervkanalnähe	
20		38		verlagert	
21		38		retiniert	
22					
23		48		Durchgebrochen	
24		48		im Durchbruch	
25		48		angelegt	
26		48		Nervkanalnähe	
27		48		verlagert	
28		48		retiniert	
29					
30					
31					
32					
33					

B		A	B	C	D	E
			Checkbox	Lokalisation:		
				rechter Ramus		
3		X		linker Ramus		
4				rechter Corpus mandibulae		
5				linker Corpus mandibulae		
6				rechte Maxilla		
7				linke Maxilla		
8						
9				Ausdehnung:		
10		X		scharf begrenzt		
11				unscharf begrenzt		
12				randständige Sklerosierungen		
13		X		mehrere Zystenräumen		
14				mit Kontinuitätsunterbrechung der Compacta		
15		X		ohne Kontinuitätsunterbrechung der Compacta		
16				eindeutige Relation zu einem Zahn		
17		X		Verdrängendes Wachstum		
18				Resorption benachbarter Zahnwurzeln		
19						
20						
21						
22						
23						
24						
25						
26						
27						
28						
29						
30						
31						
32						
33						

Abbildung 21: Die Abbildung zeigt die zwei Tabellenblätter „Sapientes“ (A) und „zystische Veränderungen“ (B), welche sich deutlich im Aufbau unterscheiden, was zu Fehlinterpretationen der KI geführt haben könnte. Es wird gezeigt, wie die einzelnen Tabellenblätter exemplarisch ausgefüllt worden sein könnten. Das „X“ markiert hierbei einen positiven Befund, der in den Röntgenbericht aufgenommen werden sollte.

9.4 Danksagung

An dieser Stelle möchte ich all jenen danken, die mich bei der Anfertigung dieser Dissertation unterstützt haben.

Ebenso danke ich allen teilnehmenden Studierenden und Patientinnen und Patienten, die durch ihre Bereitschaft zur Mitwirkung und ihren Zeitaufwand einen entscheidenden Beitrag zum Gelingen dieser Arbeit geleistet haben.

Mein besonderer Dank gilt außerdem meiner Familie: meiner Mutter und meinem Vater, die mich in meiner persönlichen und beruflichen Laufbahn stets begleitet und unterstützt haben, sowie meiner Schwester, die mir seit jeher als großes Vorbild dient.

10 Tabellarischer Lebenslauf

Name: Annika Sophie Bertsch

Berufserfahrung

01/2025 – heute Praxis für moderne Zahnheilkunde und Implantologie

Ausbildung

10/2019 – 12/2024 Studium der Zahnmedizin

Publikationen

Stephan, D., Bertsch, A., Burwinkel, M., Vinayahalingam, S., Al-Nawas, B., Kämmerer, P. W., & Thiem, D. G. (2024). AI in Dental Radiology- Improving the Efficiency of Reporting With ChatGPT: a comparative Study. *J Med Internet Res*, 26, e60684. <https://doi.org/10.2196/60684>

Stephan, D., Bertsch, A., Schumacher, S., Puladi, B., Burwinkel, M., Al-Nawas, B., Kämmerer, P. W., & Thiem, D. G. E. (2025). Artificial intelligence in dental radiology: improving patient communication with ChatGPT – a comparative study. *JMIR Preprints*, e73337. <https://doi.org/10.2196/preprints.73337>