EHEALTHCOM

MAGAZIN FÜR HEALTH-IT, VERNETZTE MEDIZINTECHNIK UND TELEMEDIZIN

www.e-health-com.de Nr. 2_3 | 2023

LLM KANN GENERATIVE KI MEHR ALS NUR STAATSEXAMEN?

DIGITALSTRATEGIE WAS STEHT DRIN UND WIE GEHT ES WEITER?

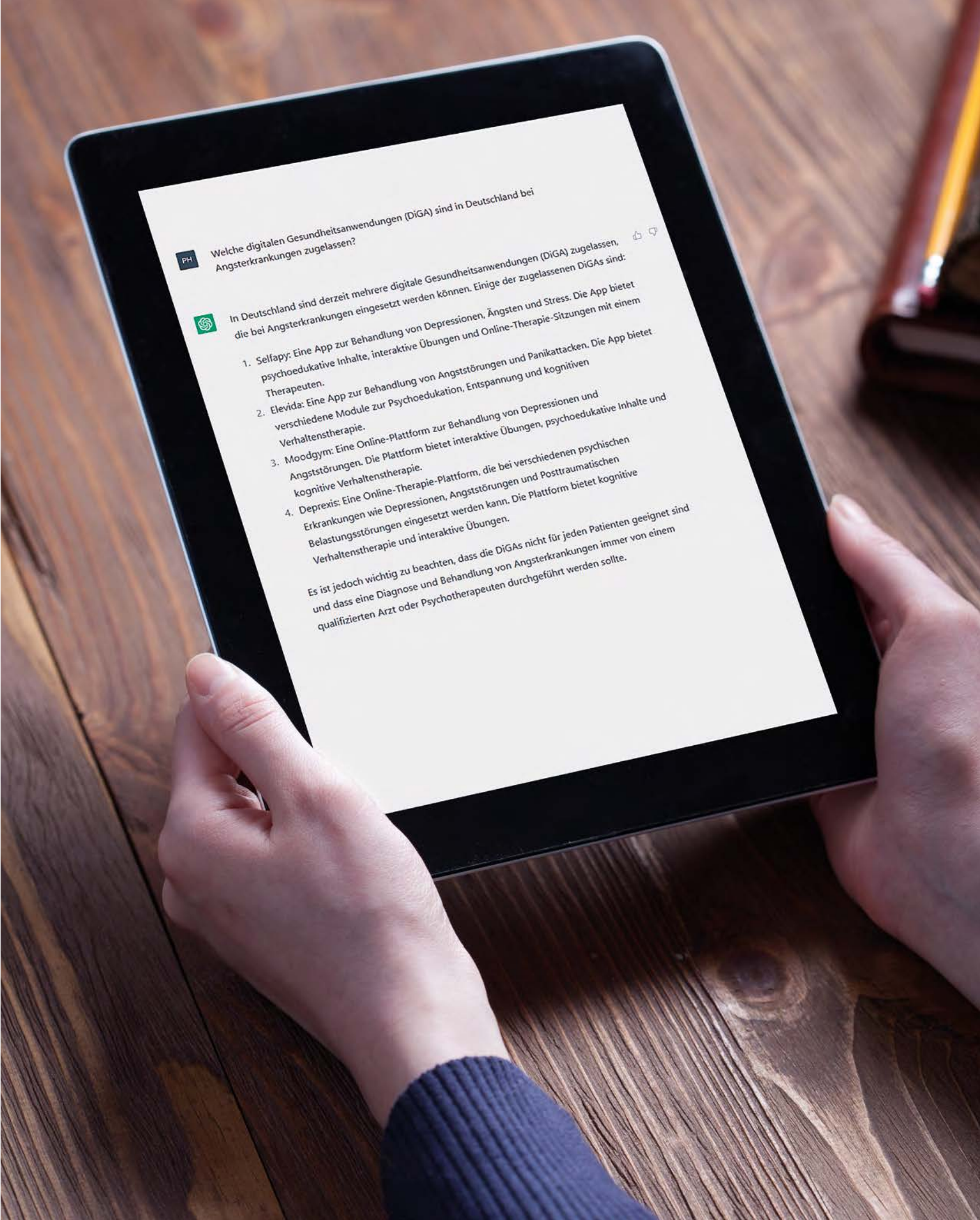
STANDARDS WANN KOMMT PLUG & PLAY IN DER MEDTECH?

DIKTIEREN SIE NOCH ODER KI-EN SIE SCHON?

Generative KI-Systeme sind derzeit in aller Munde. Selbst die Feuilletons berichten über ChatGPT, Bing-AI und Co – und natürlich gibt es auch in der Medizin einige, die die neuen, teils auch gar nicht neuen, Technologien möglichst rasch nutzen wollen. Wie realistisch ist das? Eine Analyse.

TEXT: DIRK HÜSKE-KRAUS

HL7. FHIR. DICOM.“ Das würden wohl viele auf die Frage antworten, was denn das wichtigste Kommunikations- und Dokumentationsformat in der klinischen Medizin sei. Aber zumindest so lange wie Medizin von Menschen für Menschen betrieben wird, muss die Antwort lauten: „Natürlichsprachlicher, unstrukturierter Text“. In den hochgradig arbeitsteiligen Zusammenhängen klinischer Versorgung kommunizieren die Akteur:innen in natürlicher Sprache, werden Patient:innen und Angehörige in natürlicher Sprache informiert und wesentliche Sachverhalte in natürlicher Sprache dokumentiert. Routinedokumente wie Pflegeberichte, Untersuchungsbefunde und Konsile, OP-Berichte, Herzkatheter-Protokolle, Verlegungsberichte, Arztbriefe, aber auch etwa Gutachten oder der Briefwechsel mit dem MDK: Sie alle enthalten Freitext in natürlicher Sprache, auch wenn sich hier und da eingefügte Bilder, strukturierte Befunde oder Code-Listen aus ICD oder OPS finden, und auch wenn mancherorts definierte Terminologien in radiologischen Befunden oder Pflegeberichten verwendet werden.


PH
Welche digitalen Gesundheitsanwendungen (DiGA) sind in Deutschland bei Angsterkrankungen zugelassen?
In Deutschland sind derzeit mehrere digitale Gesundheitsanwendungen (DiGA) zugelassen, die bei Angsterkrankungen eingesetzt werden können. Einige der zugelassenen DiGAs sind:
1. Selfapy: Eine App zur Behandlung von Depressionen, Ängsten und Stress. Die App bietet psychoedukative Inhalte, interaktive Übungen und Online-Therapie-Sitzungen mit einem Therapeuten.
2. Elevida: Eine App zur Behandlung von Angststörungen und Panikattacken. Die App bietet verschiedene Module zur Psychoedukation, Entspannung und kognitiven Verhaltenstherapie.
3. Moodgym: Eine Online-Plattform zur Behandlung von Depressionen und Angststörungen. Die Plattform bietet interaktive Übungen, psychoedukative Inhalte und kognitive Verhaltenstherapie.
4. Deprexis: Eine Online-Therapie-Plattform, die bei verschiedenen psychischen Erkrankungen wie Depressionen, Angststörungen und Posttraumatischen Belastungsstörungen eingesetzt werden kann. Die Plattform bietet kognitive Verhaltenstherapie und interaktive Übungen.
Es ist jedoch wichtig zu beachten, dass die DiGAs nicht für jeden Patienten geeignet sind und dass eine Diagnose und Behandlung von Angsterkrankungen immer von einem qualifizierten Arzt oder Psychotherapeuten durchgeführt werden sollte.

ANFORDERUNGEN AN „TEXTPRODUKTION" IM KRANKENHAUS

Bei der Menge an Text, die ein Krankenhaus pro Tag erzeugt, gewinnt die Frage, ob die Methoden, mit denen dieser Text „produziert" wird, ideal sind, einige Brisanz. Aber was genau heißt „ideal"? Folgende Kriterien wird man, in absteigender Reihenfolge der Wichtigkeit, anlegen wollen: Inhaltliche Korrektheit und Vollständigkeit, Lesbarkeit/Verständlichkeit, sowie Adäquatheit/formale Korrektheit und Effizienz.

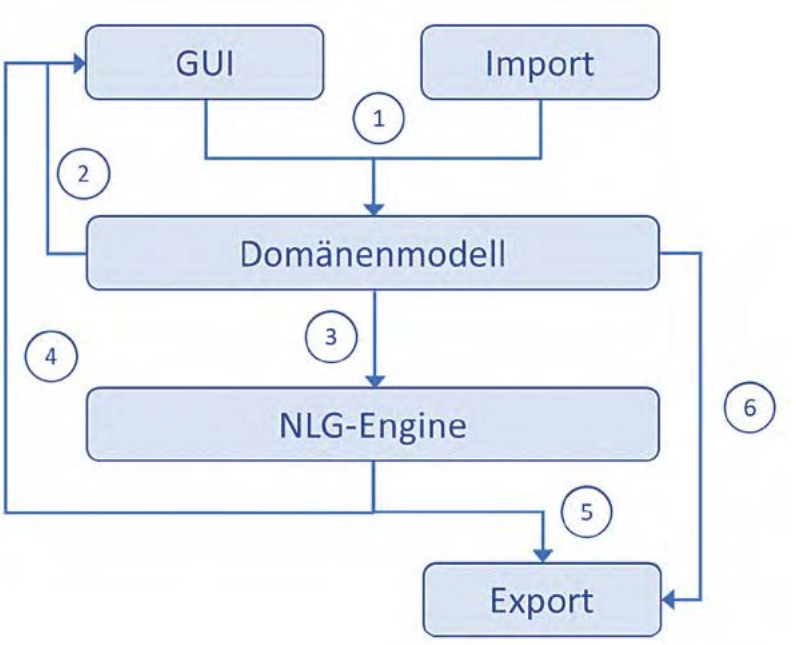
- **Korrektheit/Vollständigkeit:** So trivial es klingt, stellt es doch für einige der nachfolgend vorgestellten Verfahren eine wesentliche Hürde dar, dass alle in einem Dokument, also etwa einem OP-Bericht oder einem Entlassbrief, vorhandenen Informationen stimmen sollten und dass keine wesentlichen Informationen unter den Tisch gefallen sind.
- **Lesbarkeit/Verständlichkeit:** Natürlich sollte das Dokument auch leicht lesbar sein: Dazu gehört eine (dokumententypspezifische) Struktur, Verwendung der „richtigen" Termini, aber auch nicht ermüdende, sondern flüssige, kohärente und möglichst konzise Formulierungen. Dies kann je nach Dokumententyp auch bedeuten, dass ganze Sätze anstelle eines Telegrammstils verwendet werden, dass Fachausdrücke umschrieben oder erläutert werden (z.B. bei individualisierter Patientenaufklärung), aber auch dass bestimmte (mehrdeutige) Akronyme nicht verwendet werden sollten („HWI" kann „Harnwegsinfekt" oder „Hinterwandinfarkt" bedeuten, „MS" entweder „Mitralstenose" oder „Multiple Sklerose").
- **Adäquatheit/formale Korrektheit:** Hierzu zählt etwa die Verwendung klinikspezifischer Standards in der Strukturierung und Terminologie, Berücksichtigung der Empfänger und deren sprachlicher Kompetenzen, aber auch scheinbar triviale Themen wie Orthografie und Grammatik. Gerade für Nicht-Muttersprachler:innen sind hier oft selbst banal erscheinende Qualitätsanforderungen nicht leicht zu erfüllen.
- **Effizienz:** Das Erstellen klinischer Dokumente ist in der Regel eine ungeliebte Tätigkeit. Niemand wird Chirurg:in, weil er oder sie so gerne OP-Berichte schreibt. Das Zusammensuchen und Abdiktieren von Befundteilen aus der Fallakte bei der Erstellung des Arztbriefes ist nicht nur fehleranfällig, sondern auch zeitaufwendig. Die ärztliche Arbeitskraft ist eine viel zu gesuchte Ressource, als dass man sie auf redaktionelle Tätigkeiten, das Navigieren durch labyrinthische Textbausteinverzeichnisse oder das hundertfach wiederholte Diktieren nahezu identischer Textblöcke verschwenden sollte.

Legt man diese Kriterien an, dann sieht man, dass „klassische" Verfahren der Produktion klinisch relevanter Texte oft defizitär sind:

1. Selbst tippen: Auch bei Anwender:innen, die perfekt Zehn-Finger-Schreiben können, ist dieses Verfahren sehr aufwendig. Es gibt keine Vollständigkeits- oder Plausibilitätskontrolle, die Dokumentenstruktur muss jedes Mal neu erzeugt werden, und natürlich ist neben nicht entdeckten Tippfehlern hier das Kostenargument einschlägig.
2. Diktieren und Tippen lassen: Es ist erstaunlich, wie häufig dieses Verfahren (auch mit physischem Transport von Magnetbandkassetten) noch verwendet wird. Die Ressource „Arzt/Ärztin" wird hier zwar nicht für das Tippen ver(sch)-wendet, dafür aber eine weitere Person, und der Diktat-Tipp-Korrekturzyklus wird oft mehrfach durchlaufen. Und auch hier gilt, dass es keine automatisierte Qualitätskontrolle gibt und dass – wie auch bei 1. – Nicht-Muttersprachler:innen vor besonderen Schwierigkeiten stehen.
3. Letzteres gilt auch für den natürlich erscheinenden Verbesserungsschritt: Die automatisierte Spracherkennung. Auch hier wird der Diktierende zum Korrektor der nicht selten absonderlichen Einfälle des Spracherkenners, und es gibt – jenseits der nächsten Stufe in der Vidierungspipeline – keine Plausibilitäts-, Vollständigkeits- oder auch Adäquatheitskontrolle. Für Texte, die sich oft nur in wenigen Details unterscheiden, etwa OP-Berichte nach Standardoperationen, ist dieser Ansatz auch ineffizient, da auch die sich wiederholenden Passagen jedes Mal neu diktiert werden müssen.
4. Für einige Anwendungen sind Templates, also Lückentexte mit Variablen, oder Textbausteine eine brauchbare Lösung. Aber schon bei mäßiger Komplexität begrenzen die mangelnde Kombinierbarkeit von Textbausteinen, die Notwendigkeit von Flexionsformen („der Patient", „des Patienten", „der Patientin") und die Starrheit des Templates die Ausdrucksfähigkeit übermäßig.

WAS TUN?

Sprachtechnologie gewinnt einen immer größeren Raum und eröffnet sich immer mehr Anwendungsbereiche. In der Medizin sind dies hauptsächlich Verfahren der „Natural Language Analysis" (NLA), die Freitexte analysieren und daraus strukturierte Informationen gewinnen können. Die Verschlüsselung von Diagnosen aus Klartext ist hier ein bekanntes Beispiel. Systeme wie DaWiMed [1] können aus den freitextlichen Einträgen der medizinischen Dokumentation strukturier-



SCHEMATISCHER ABLAUF DER DOKUMENTERSTELLUNG IN B|FLOW: Aus den importierten und im GUI eingegebenen Daten (1) wird das Domänenmodell befüllt (1). Dieses kann Plausibilitäts- und Vollständigkeitsprobleme im GUI anzeigen oder korrigieren (2). Auf Basis des Domänenmodells (3) erzeugt die NLG-Engine den Zieltext und zeigt diesen im GUI an (4). Der Zieltext kann bei Abschluss an das KIS/KAS übergeben werden (5), ebenso wie strukturierte Informationen aus dem Domänenmodell (6).

te Informationen extrahieren, die dann etwa für Qualitätskontrolle, Studienrekrutierung oder auch als „watchdog" für epidemiologische Fragestellungen oder Orphan Diseases verwendet werden können.

Weniger präsent in der Medizin ist der zweite „Arm" von Sprachtechnologie, nämlich „Natural Language Generation" (NLG), also die Erzeugung natürlicher Sprache aus strukturierter Information. NLG adressiert genau das oben beschriebene Problem der Produktion klinischer Texte.

Kommerziell werden NLG-Systeme heute in vielen nichtmedizinischen Kontexten eingesetzt: Niemand erzeugt die Texte in der App für die Wettervorhersage in Klein-Kleckersdorf von Hand, und Betreiber von großen Online-Shops lassen die Artikelbeschreibungen aus strukturierten Daten von NLG-Systemen erstellen. Ähnliches gilt etwa auch für Exposés für Objekte auf Immobilienseiten. Sportberichte für niederklassige Begegnungen werden oft schon durch „Robo-Journalists" [2] geschrieben, und in den USA arbeitet ArriaNLG an einer voll automatisierten Liveberichterstattung für etwa Basketballspiele [3, 4]. Zum Glück ist Béla Réthy schon in Rente.

Aktuell gibt es eine große Aufmerksamkeit für ChatGPT [5], und die Ergebnisse dieses Systems verblüffen tatsächlich.

WARUM SCHREIBT NICHT CHATGPT LÄNGST DIE BEFUNDE?

Unlängst ging ein Video viral, in dem ein US-amerikanischer Rheumatologe beschreibt, wie ChatGPT für ihn die Briefe schreibt, in denen er die Notwendigkeit bestimmter Untersuchungen begründet, inklusive Verweise auf die einschlägige Literatur. Nach Aussage des Arztes eine fantastische Zeitersparnis: „Amazing stuff. Use this in your daily practice, okay. It will save time. It will save effort." [6]

Dabei ist das Feld „NLG" nicht brandneu: Erste Publikationen finden sich schon in den 1980ern, im Jahr 2003 erschien der erste Review-Artikel zu „Text Generation in Clinical Medicine" [7], und das System BabyTalk etwa [8] erzeugte im Jahr 2008 bereits Übergabeberichte für eine neonatologische Intensivstation. Zwar wurde bislang keines der im Healthcare-Umfeld angesiedelten Systeme je kommerziell eingesetzt, aber man könnte vermuten, dass dieser Schritt heute kurz bevorsteht.

Ehud Reiter, einer der Väter von BabyTalk und über Dekaden einflussreich im NLG-Bereich, ist kritisch. In seinem Blog [9] benennt er als ein Hauptproblem die mangelnde Korrektheit der generierten Texte, insbesondere die Neigung von GPT zu „Halluzinationen". Das System erfindet Dinge, die sehr plausibel klingen, aber leider nicht wahr sind. So auch im obigen Beispiel des Rheumatologen: Der in dem TikTok-Video zu sehende Brief enthält zwei Referenzen auf Studien, die aber gar nicht existieren. Mittlerweile sieht der Arzt die Angelegenheit deutlich zurückhaltender: „So it's not something you could just send off to an insurance company, or really use at this point." [10]

Während die frühen NLG-Systeme deterministisch waren, also vordefinierten Abläufen folgten, basieren moderne Systeme wie GPT4 wesentlich auf probabilistischen Verfahren, gestützt auf große Mengen an Trainingsdaten. Durch diesen probabilistischen Ansatz und die damit einhergehende Tendenz zu Konfabulationen >

Die Systeme, die gerade in aller Munde sind, wie ChatGPT, scheinen den speziellen Anforderungen klinischer Dokumentenerzeugung nicht gewachsen.

wird die Einsetzbarkeit dieser Systeme im ganz spezifischen Umfeld klinischer Dokumente aber fragwürdig. Die Neigung zu Halluzinationen disqualifiziert eigentlich jeden Befundersteller oder Arztbriefschreiber. Ein weiteres Problem benennt Reiter: die Evaluation. Wie soll man systematisch bewerten, ob der Output den gewünschten Qualitätskriterien entspricht? Herkömmliche Metriken für die Bewertungen von Textqualität und -verständlichkeit (siehe etwa [11] für einen Überblick) sind nicht unmittelbar zu verwenden.

Und schließlich braucht man, um auch nur annähernd korrekte Texte zu erzeugen, unrealistisch große Mengen an annotierten Trainingsdaten. Bereits einfache technische Untersuchungen haben schnell 15-25 Parameter. Nimmt man der Einfachheit halber an, dass diese Parameter alle boolesch sind, dann kommt man bereits auf zwischen 32 000 und 33 Millionen annotierter unterschiedlicher Befunde, mit denen man das System trainieren müsste. „Annotiert" ist hier das Schlüsselwort, denn auch wenn man etwa 33 Millionen Ultraschallbefunde des Abdomens fände, ist es doch nahezu unmöglich, diese zu annotieren, also die strukturierte Information beizufügen, welche Einzelbeobachtungen vorlagen und welche nicht.

Weitere Gründe, warum ChatGPT und verwandte Systeme für die Erzeugung klinischer Dokumente ungeeignet sind, nennt etwa [12]. So aufregend etwa die Erfahrungen mit GPT sind, der Weg in die Erstellung von klinischen Routinedokumenten mit derartigen Systemen scheint (noch?) nicht offen.

ANDERE ANSÄTZE: CHATGPT IST NICHT ALLES

Aber es gibt andere Ansätze. All diesen ist gemein, dass die jeweilige Anwendungsdomäne durch menschliche Experten modelliert und relevante Sachverhalte in ihrer textuellen Beschreibung in einem gewissen Maße vordefiniert werden. So setzen etwa bereits in Deutschland verfügbare Befundungssysteme auf Formularbibliotheken, mit denen die Anwender ihre Texte durch Eingaben in einem grafischen Interface (GUI) erstellen

können. Strukturierte Daten, die ansonsten durch NLA aus den Freitexten extrahiert werden müssen, fallen bei einem derartigen Ansatz gewissermaßen als Nebenprodukt mit ab. Während diese Systeme in der Regel mit Textbausteinen arbeiten und somit auch deren unvermeidlichen Beschränkungen unterliegen, geht das System MARIS B|flow [13] einen Schritt weiter: Hier sorgt eine computerlinguistische Komponente dafür, dass stets syntaktisch korrekte Texte erzeugt werden. Auch z. B. das Problem von Aggregation, also dem Zusammenfassen von Einzelbeobachtungen zu einem Gesamt-(Fach-)ausdruck lässt sich so lösen.

Der Ansatz, Eingaben in ein GUI mit einem Modell der Anwendungsdomäne und einer, möglichst auf computerlinguistischen Verfahren basierenden, Komponente der Textgenerierung zu kombinieren (siehe Infokasten), hilft insgesamt, die oben eingeführten Qualitätskriterien mindestens teilweise zu erfüllen:

- Vollständigkeits- und Plausibilitätsbedingungen können im Domänenmodell hinterlegt und im GUI signalisiert werden.
- Auch Nicht-Muttersprachler können so korrekte Texte in der Zielsprache erzeugen.
- Die Befundformulare können so gestaltet werden, dass die hausinternen Formanforderungen und fachspezifische Inhaltsanforderungen erfüllt werden. Darüber hinaus hat ein GUI auch noch eine edukative Funktion: Eher unerfahrene Befundersteller werden so durch alle Bereiche geleitet, zu denen sie Stellung nehmen sollten.

Einen weiteren Vorteil bietet NLG gegenüber herkömmlichen Verfahren der Dokumenterstellung wie Diktat oder Templates. Es lassen sich nämlich, mindestens prinzipiell, ohne Mehraufwand bei der Dokumenterstellung verschiedene Versionen des Dokumentes erzeugen. Etwa eine weitere Version in patientengerechter Sprache oder auch eine zusätzliche Version beispielsweise in Englisch oder Italienisch. In mehrsprachigen Ländern, aber auch für Einrichtungen mit einem hohen Anteil an ausländischen Patienten, lässt sich so eine Serviceverbesserung ohne Übersetzungskosten erzielen.

FAZIT

Insgesamt liefert der Ausblick auf NLG-Einsatz in der Medizin ein zwiespältiges Bild: Die Systeme, die gerade in aller Munde sind, wie ChatGPT, scheinen den speziellen Anforderungen klinischer Dokumentenerzeugung nicht gewachsen. Die Alternative besteht in Systemen, die die Anwendungsdomäne modellieren und aus GUI-Eingaben (oder importierten Daten) mittels computerlinguistischer Verfahren einen syntaktisch korrekten, semantisch plausiblen und pragmatisch adäquaten Text erzeugen. Selbst in einer syntaktisch eher anspruchsvollen Sprache wie dem Deutschen zeigen diese Systeme einen gangbaren und zielführenden Weg auf. ■

■ **DIRK HÜSKE-KRAUS**
hat über NLG in der Medizin promoviert und ist heute selbstständig in der IT-Beratung im Gesundheitswesen.
Kontakt: dirk@suregen.de

QUELLEN

1. ID GmbH, 2022: Die eigenen Daten besser kennenlernen, https://www.id-berlin.de/magazin-post/die-eigenen-daten-besser-kennenlernen-freitext-scannen-mit-hilfe-von-terminologien-und-fhir/, zuletzt besucht 27.2.2023

2. Retresco: Was ist Roboterjournalismus?, https://www.retresco.de/ressourcen/lexikon/lexikoneintrag/roboterjournalismus, zuletzt besucht 27.2.2023

3. Presseportal.de, 2021: Arria NLG erwirbt Boost Sport AI, https://www.presseportal.de/pm/133435/5087930, zuletzt besucht 27.2.2023

4. Arria: Boost Sport AI, https://www.arria.com/sports/, zuletzt besucht 27.2.2023

5. OpenAI: ChatGPT, https://chat.openai.com/chat, zuletzt besucht 27.2.2023

6. Stermer C, 2022, https://www.tiktok.com/@tiktokrheumdok/video/7176660771806383403, zuletzt besucht 27.2.2023

7. Hüske-Kraus D., 2003: Text generation in clinical medicine – a review. Methods of information in medicine, 42(01), 51-60.

8. Reiter E, et al., 2008: The importance of narrative and other lessons from an evaluation of an NLG system that summarises clinical data. Proceedings of the INLG-08.

9. Reiter E: Ehud Reiter's Blog, https://ehudreiter.com/, zuletzt besucht 27.2.2023

10. MedPage Today, 29.12.2022, What Can ChatGPT Do For Your Practice?, https://www.medpagetoday.com/special-reports/exclusives/102312, zuletzt besucht 27.2.2023

11. Kiefer C, 2019: Quality indicators for text data. BTW 2019–Workshopband, verfügbar über https://btw.informatik.uni-rostock.de/download/workshopband/C2-5.pdf, zuletzt besucht 27.2.2023

12. John Snow Labs, An early evaluation of ChatGPT on common medical NLP tasks, https://www.johnsnowlabs.com/an-early-evaluation-of-chatgpt-on-common-medical-nlp-tasks/, zuletzt besucht 27.2.2023

13. MARIS: MARIS B|flow, https://maris-healthcare.de/produkte/maris-befundbrief/nlg/, zuletzt besucht 27.2.2023