Automatische Beantwortung von Gesundheitsfragen durch generative KI mittels Retrieval Augmented Generation

Vergleich und Potentiale von Großen Sprachmodellen anhand des Beispiels Diabetes mellitus Typ 2

S. Lengauer[1] F. Proprentner[1] M. Tytarenko[1] C. Krenn[2] K. Jeitler[2,3] T. Schreck [1]

[1]Technische Universität Graz, Institut für Computer Graphik und Wissensvisualisierung
[2]Medizinische Universität Graz, Institut für Allgemeinmedizin und evidenzbasierte Versorgungsforschung
[3]Medizinische Universität Graz, Institut für Medizinische Informatik, Statistik und Dokumentation

Hintergrund

Die Fortschritte in der Entwicklung von KI-gestützten großen Sprachmodellen (LLMs) haben diese in vielen Bereichen zu leistungsfähigen Werkzeugen gemacht, die menschliche Entscheidungsträger zunehmend automatisiert unterstützen. LLMs wie GPT, GEMINI, LLAMA oder MISTRAL werden mit allgemeinen Textdaten trainiert und sind in der Lage, natürlichsprachliche Texte zu erzeugen, die oft kaum von Expert*innentexten zu unterscheiden sind. Im Bereich der Gesundheitsinformationssysteme können LLMs z.B. in Form von Chatbots eingesetzt werden, um auf spezifische Informationsbedürfnisse von User*innen individuell einzugehen. Dies stellt jedoch hohe Anforderungen an die Qualität der Antworten hinsichtlich inhaltlicher Korrektheit, Relevanz und Angemessenheit.

Eine der größten Herausforderungen beim Einsatz von LLMs in Gesundheitsinformationssystemen ist der Umgang mit KI-Halluzinationen – plausibel erscheinende Informationen, die aber inhaltlich falsch sind. Um die Zuverlässigkeit der Antworten auf Gesundheitsfragen zu erhöhen, kann die Methode der Retrieval Augmented Generation (RAG) eingesetzt werden. Durch Kombination eines LLMs mit zuverlässigen externen Informationsquellen als Basis für die Antworten soll die Robustheit der generierten Antworten erhöht und das Auftreten von KI-Halluzinationen reduziert werden.

Methodik

In unserer Studie verglichen wir eine Auswahl mehrerer verbreiteter LLMs hinsichtlich ihrer Fähigkeit, Gesundheitsfragen zu Diabetes Typ 2 mithilfe der RAG-Methode zu beantworten. Dafür wurde folgender Prompt verwendet:

> *Verhalte dich wie ein evidenzbasierter klinischer Forscher. Antworte kurz und bündig. Ist die Antwort nicht in den gegebenen Fakten enthalten, antworte mit ‚Das ist nicht in meinen Informationen enthalten'. Beantworte bitte die folgende Frage (<FRAGE>) ausschließlich mit den folgenden Fakten: (<CONTEXT>).*

Als Informationsquelle (also 'Context') diente eine etablierte Patienteninformationsbroschüre [1]. Wir definierten einen Benchmark mit häufig gestellten Gesundheitsfragen zu Diabetes Typ 2 und qualitätsgesicherten Referenzantworten. In quantitativen Experimenten bewerteten wir die von den LLMs generierten Antworten mit etablierten Metriken zur Berechnung der Textähnlichkeit im Vergleich zur Referenz.

Modelle

GEMINI 1.5 (Google) `https://deepmind.google/technologies/gemini/`
GPT 3.5/4o (OpenAI) `https://openai.com/index/hello-gpt-4o/`
LLAMA3 (Meta) `https://www.llama.com/`
MISTRAL (Mistral AI) `https://mistral.ai/`

Benchmark Fragen

Mit der Auswahl an Fragen versuchten wir, verschiedene Aspekte von Fragen – z.B., Offene versus Geschlossene, Faktenwissen versus Interpretation – abzudecken.

1. Was ist Diabetes?
2. Was sind normale Werte für den Blutzuckerspiegel?
3. Was ist der HbA1c-Wert?
4. Was ist der Unterschied zwischen Typ 1 und Typ 2 Diabetes?
5. Welche Faktoren erhöhen das Risiko für Diabetes?
6. Welche Rolle spielt die Ernährung bei Diabetes?
7. Ist ein Wert von 500 mg/dl normal?
8. Wie kann ich vorbeugen?

Alle davon können mit den Informationen aus der genannten Broschüre [1] zufriedenstellend beantwortet werden. Für jede Frage erstellten wir auch eine entsprechende *Ground Truth* Antwort aus ebendieser. Um Varianzen in der Beantwortung zu berücksichtigen, wurde jede Frage 20 Mal pro LLM gestellt.

Evaluierungsmetriken

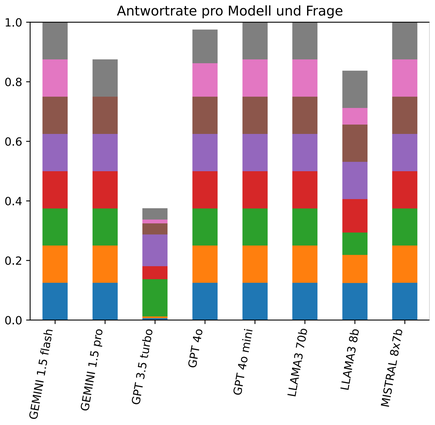
ROGUE-L Maß für die Häufigkeit gleicher Wortfolgen, basierend auf der längsten gemeinsamen Teilfolge [2].

BERT-Score Semantischer Vergleich durch Bidirectional Encoder Representation from Transformers (BERT) Embeddings.

Relevancy Genauigkeit mit der die Frage von der erhaltenen Antwort rekonstruiert werden kann [3].

Correctness Vergleich durch faktische Ähnlichkeit der Kernaussagen [3].

Ergebnisse



In einem ersten Schritt überprüften wir die Fähigkeit der Modelle, Fragen anhand der zur Verfügung gestellten Informationen grundsätzlich zu beantworten. Dafür wurden die Antworten mit der, im Prompt angegebenen Standardantwort verglichen und die Anzahl der "Nicht-Antworten" pro Frage gezählt (Abbildung links).

Anschließend wurden valide Antworten mit den angegebenen Metriken im Bezug auf die Ground Truth überprüft (Tabelle 1).

Tabelle 1. Die Modelleigenschaften und Evaluierungsergebnisse (Mittelwert/Median) gemäß der verwendeten Metriken.

Model				Scores			
Name	*Context†*	*Input‡ $*	*Output‡ $*	*ROUGE-L*	*BERT*	*Corr.*	*Relevancy*
GEMINI 1.5 flash	128	**0.08**	**0.30**	.18/.17	.70/.68	.55/.61	.92/.93
GEMINI 1.5 pro	128	3.50	10.50	**.24**/.18	**.71**/**.70**	**.67**/.57	.94/.95
GPT-3.5 turbo	16	0.50	1.50	.20/.19	.70/.69	.55/.60	.91/.95
GPT-4o	128	5.00	15.00	.20/.22	.69/.68	.60/.61	.80/.96
GPT-4o mini	128	0.15	0.60	.20/**.22**	.67/.69	.66/**.73**	.94/.95
LLAMA3 70b*	8	0.65	2.75	.19/.19	.69/.68	.59/.67	**.95**/**.96**
LLAMA3 8b*	8	**0.05**	**0.25**	.19/.19	.69/.68	.51/.51	.64/.94
MISTRAL 8x7b*	32	0.30	1.00	.16/.18	.68/.68	.64/.69	.92/.94

† Größe des Kontextfensters in K-Token.
‡ Preis pro 1M Tokens in Dollar.
* Open Source Modelle. Die angegebenen Preise beziehen sich auf Preise der verwendeten Drittanbieter (können je nach Anbieter variieren).

Schlussfolgerungen

- Die meisten Modelle lieferten kontinuierlich valide Antworten auf alle Fragen.
  - Gelegentlich invalide Antworten von *GPT-4o* und *LLAMA3 8b* (keine Muster erkennbar).
  - *GEMINI 1.5 pro* retournierte immer valide Antworten, außer auf Frage 7.
  - *GPT-3.5 turbo* konnte nur 37.5% der Fragen erfolgreich beantworten.
- *GEMINI 1.5 pro* erzielte die besten Ergebnisse.
- *GPT-4o mini* erzielte die beste Preis-Leistung – es wurden ähnliche bzw. sogar marginal bessere Ergebnisse, als von dem wesentlich größere *GPT-4o*, retourniert.

Große Sprachmodelle entwickeln sich schnell weiter. Unsere Studie bietet eine aktuelle Momentaufnahme in einem spezifischen Bereich der Gesundheitsinformationen, und deutet auf ein vielversprechendes Potential hin. Die verwendete Vergleichsmethodik ist auf zukünftige Modelle übertragbar.

References

[1] J. Baumgart, U. Viegener, and C. Pohl. Den Diabetes im Griff: Ein Handbuch für Patientinnen und Patienten mit Diabetes mellitus Typ 2. *AOK-Bundesverband, Berlin*, 2021. URL `https://www.aok.de/pk/magazin/cms/fileadmin/pk/pdf/patientenhandbuch-diabetes.pdf`.

[2] Chin-Yew Lin. ROGUE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013`.

[3] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, 2024.