INSTITUTE OF VISUAL COMPUTING

TU Graz
SCIENCE PASSION TECHNOLOGY

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], und T. Schreck[1]

[1] Technische Universität Graz, Institute of Visual Computing
[2] Medizinische Universität Graz, Institut für Allgemeinmedizin und evidenzbasierte Versorgungsforschung
[3] Medizinische Universität Graz, Institut für Medizinische Informatik, Statistik und Dokumentation

Hintergrund

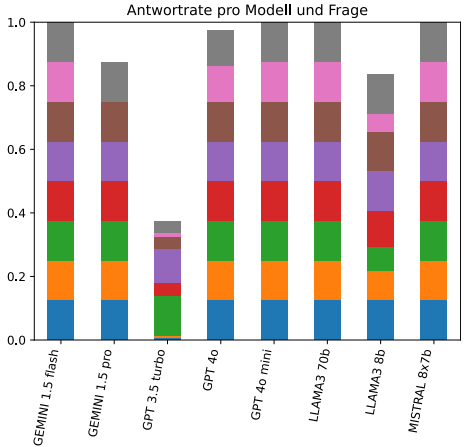
- Wachsende Relevanz von großen Sprachmodellen (LLMs) für die Vermittelung von Gesundheitsinformation
- **Herausforderung**: KI-Halluzinationen – plausibel erscheinende, jedoch **inhaltlich falsche** Informationen
- Abhilfe durch Methode der **Retrieval Augmented Generation** (RAG)
  - LLM in Verbindung mit zuverlässigen externen Informationsquellen
- Ziel unserer Studie
  - Evaluierung der RAG Methode hinsichtlich der Beantwortung von Gesundheitsfragen zu Diabetes mellitus Typ 2
  - Vergleich verschiedener moderner LLMs

Methodik

- Wiederholte Befragung der Modelle mit einem Set von Benchmark Fragen
- Auswertung der Antworten bzgl. qualitätsgesicherter Referenzantworten mit verschiedenen Evaluierungsmetriken

- Modelle
  - GEMINI 1.5 (Google)
  - GPT 3.5/4o (OpenAI)
  - LLAMA3 (Meta)
  - MISTRAL (Mistral AI)]
- Benchmark Fragen
  1. Was ist Diabetes?
  2. Was sind normale Werte für den Blutzuckerspiegel?
  3. ...

Ergebnisse



- Manche Modelle sind nicht in der Lage gewisse Fragen zu beantworten

Ergebnisse

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).