UNIVERSITÄT
LEIPZIG

Medizinische Fakultät

Abschlussvortrag

Question Answering auf dem Lehrbuch „Health Information Systems“ mit Hilfe von unüberwachtem Training eines Pretrained Transformers

Leipzig, November 2023

Paul Keller

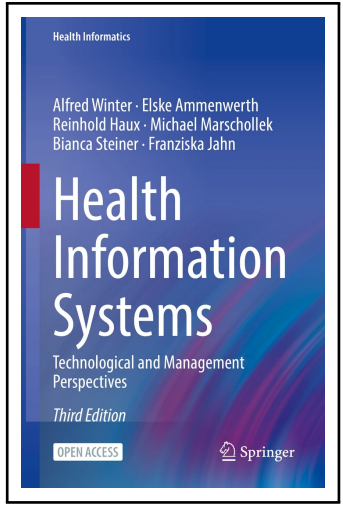
imise.

Überblick

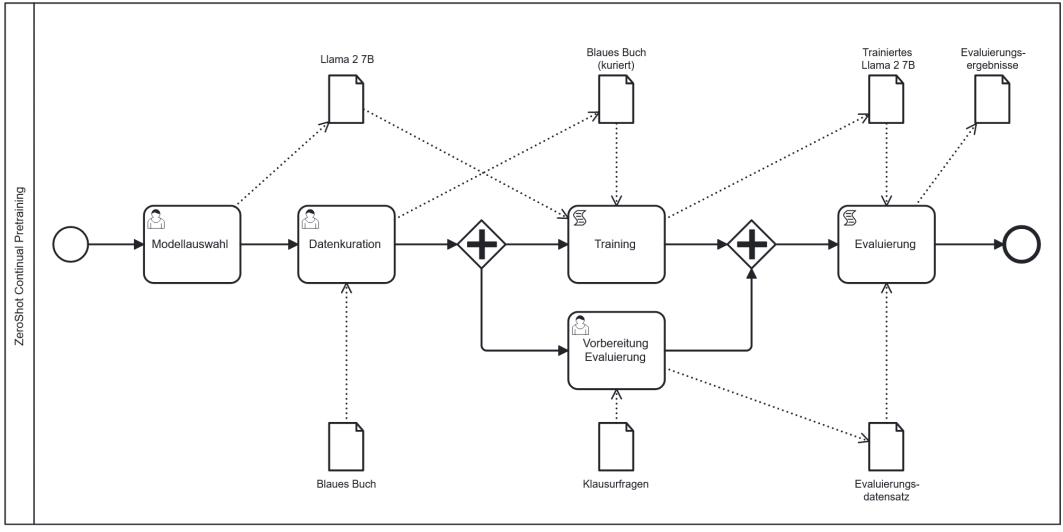
- Einleitung
  - Ziel
  - Aufgaben
- Ergebnisse
  - Evaluierte Modelle
  - Kriterien
  - Vergleich
- Diskussion und Ausblick

Einleitung - Ziel

- Fragen beantworten zu „Health Information Systems“ mit eines Sprachmodells
- Beispielklausuren lösen aus dem Modul „Architektur von Informationssystemen im Gesundheitswesen“ mit eines Sprachmodells



Einleitung - Aufgaben



Ergebnisse - Evaluierte Modelle

Modell	Epochen	Grafikkarten	Bezeichnung
GPT4	-	-	`gpt4`
Llama 2 7B	0	Nvidia Tesla A100	`llama2_0e`
	1	Nvidia Tesla V100	`llama2_1e_v100`
	1	Nvidia Tesla A30	`llama2_1e_a30`
	3	Nvidia Tesla V100	`llama2_3e_v100`
	3	Nvidia Tesla A30	`llama2_3e_a30`
	5	Nvidia Tesla A30	`llama2_5e_a30`
	10	Nvidia Tesla A30	`llama2_10e_a30`

Tabelle: Evaluierte Modelle, deren Anzahl an Epochen, genutzte Grafikkarten und Bezeichnungen in den Grafiken.

Ergebnisse - Kriterien

Aus Methodik übernommen (Omar et al.):

- Korrektheit (MakroF1)
- Erklärbarkeit
- Fragenverständnis
- Robustheit

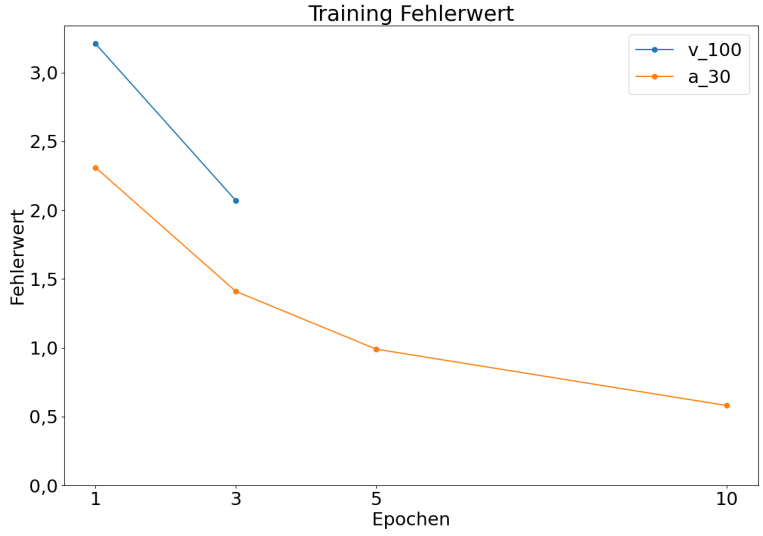
Zusätzlich gezeigt:

- Fehlerwerte (Training, Validierung)

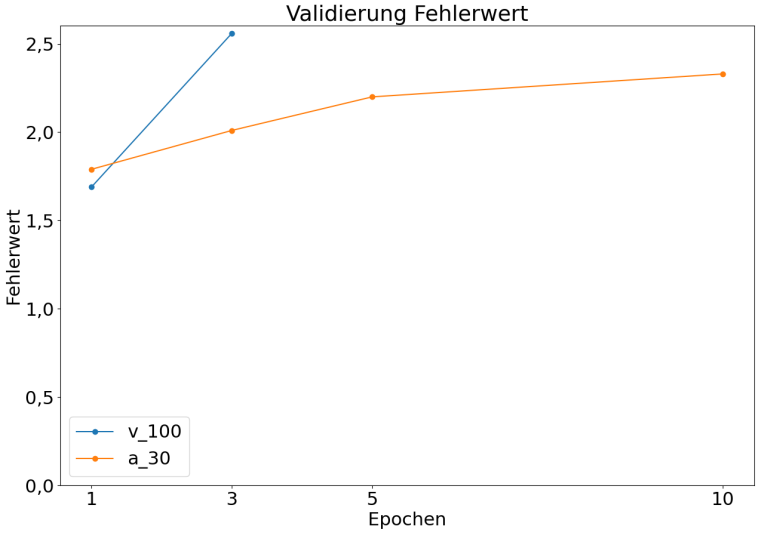
Ausgeschlossen:

- Determinismus

Kriterium Fehlerwerte



Kriterium Fehlerwerte



Evaluierungsdatensatz

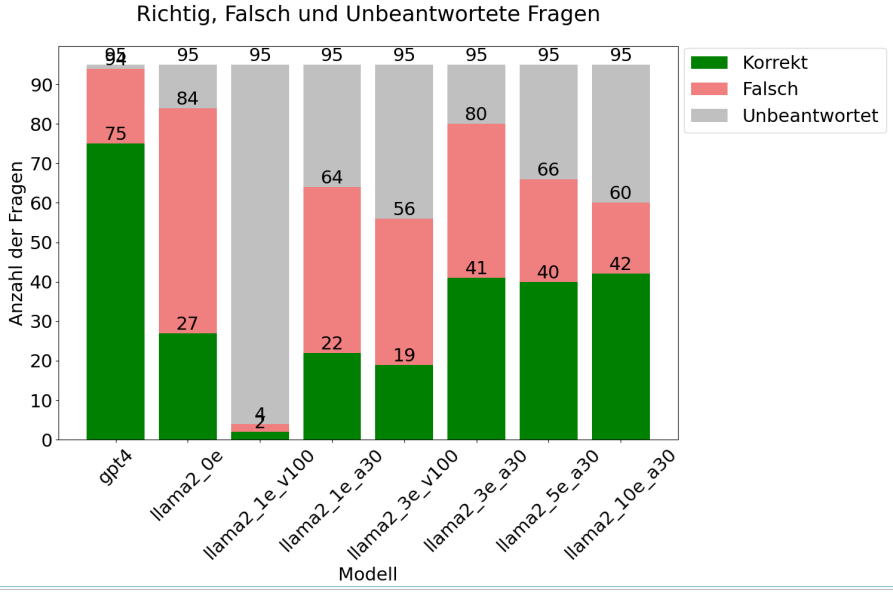
Fragen nach Fragetyp

Einzelfakt	34
Multifakten	38
Transfer	23

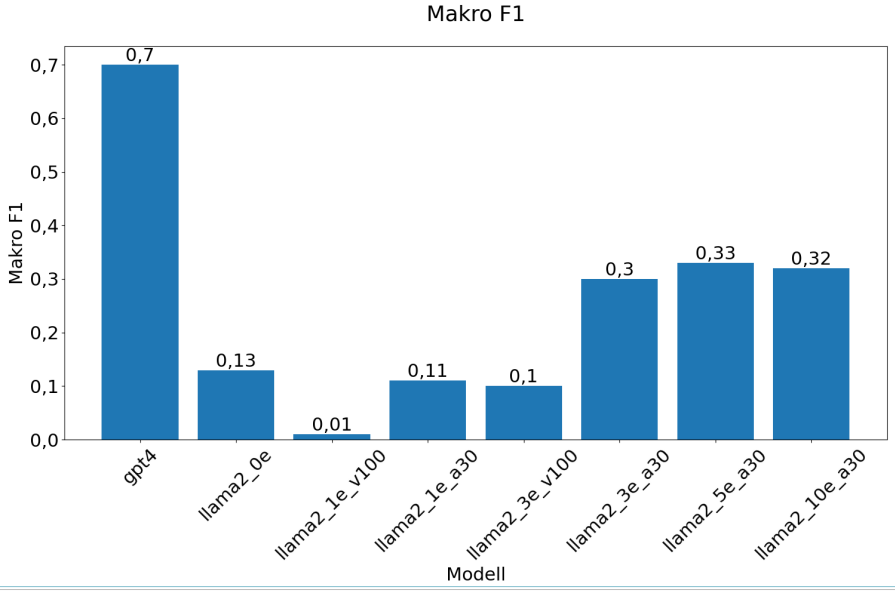
Fragen nach Quelle

Aus „Health Information Systems“	33
Informationssysteme in der medizinischen Versorgung und Forschung - Klausur	9
Informationssysteme in der medizinischen Versorgung und Forschung - Nachholklausur	31
Architektur von Informationssystemen im Gesunheitswesen - Klausur	22

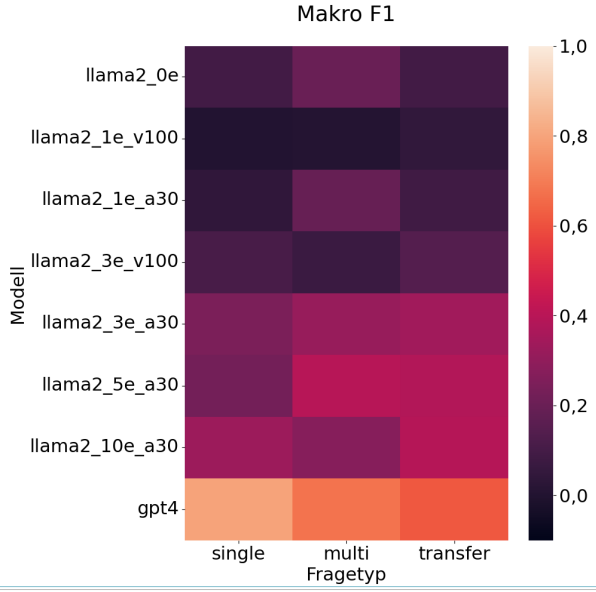
Kriterium Korrektheit



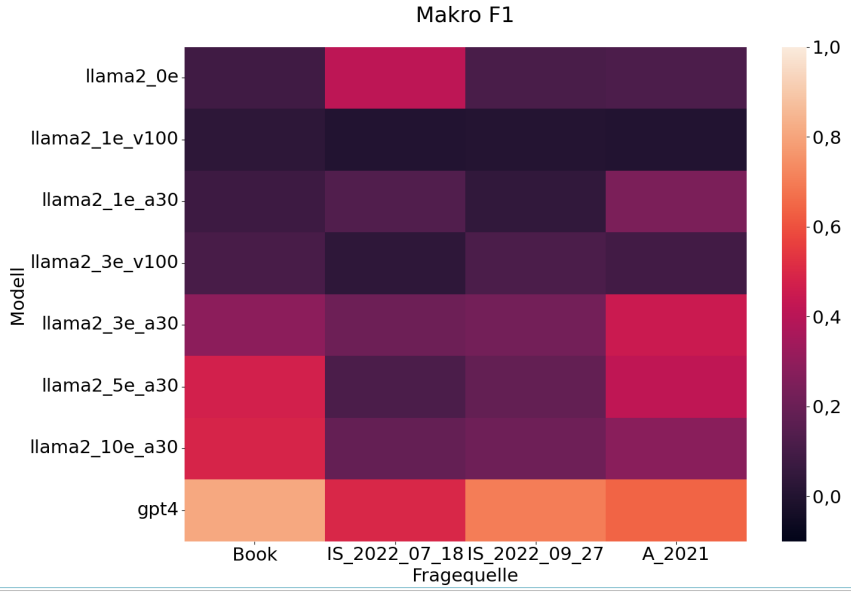
Kriterium Korrektheit



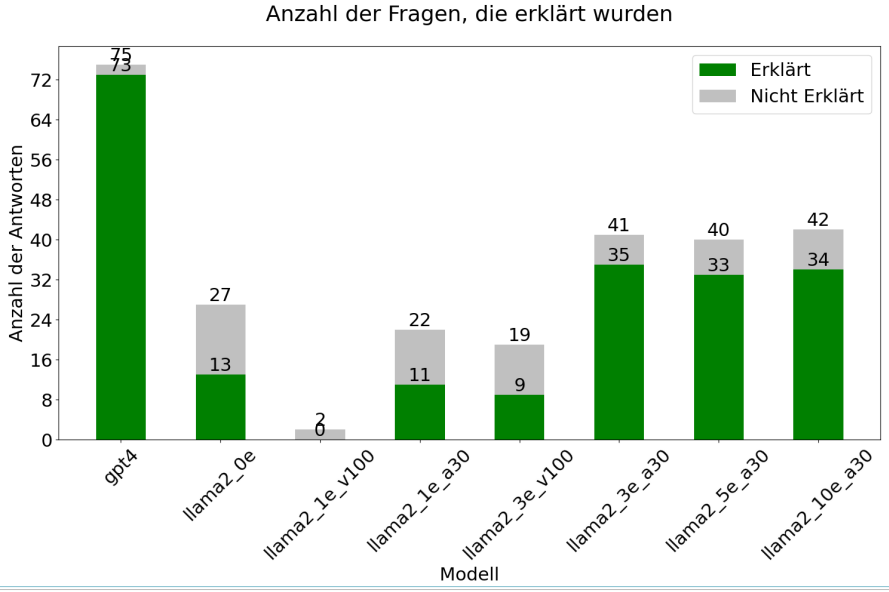
Kriterium Korrektheit



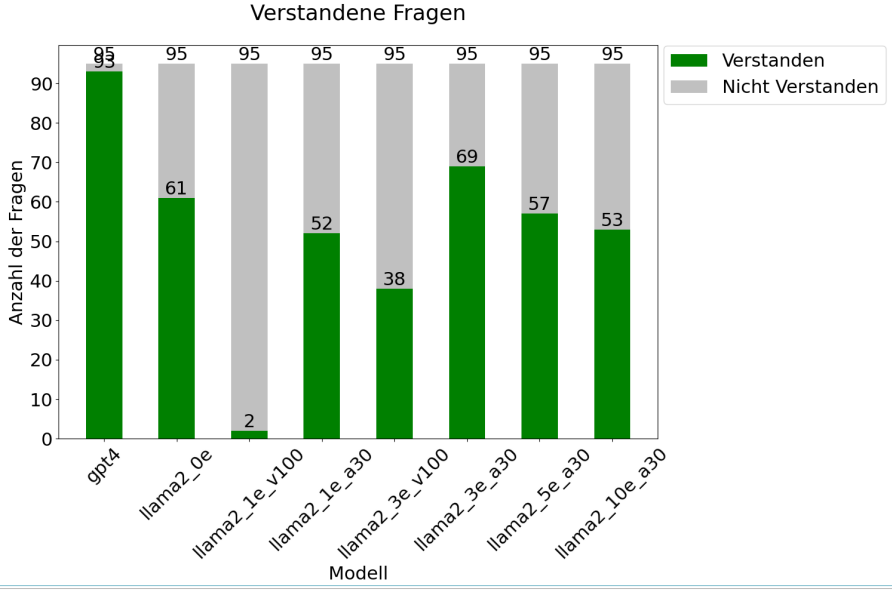
Kriterium Korrektheit



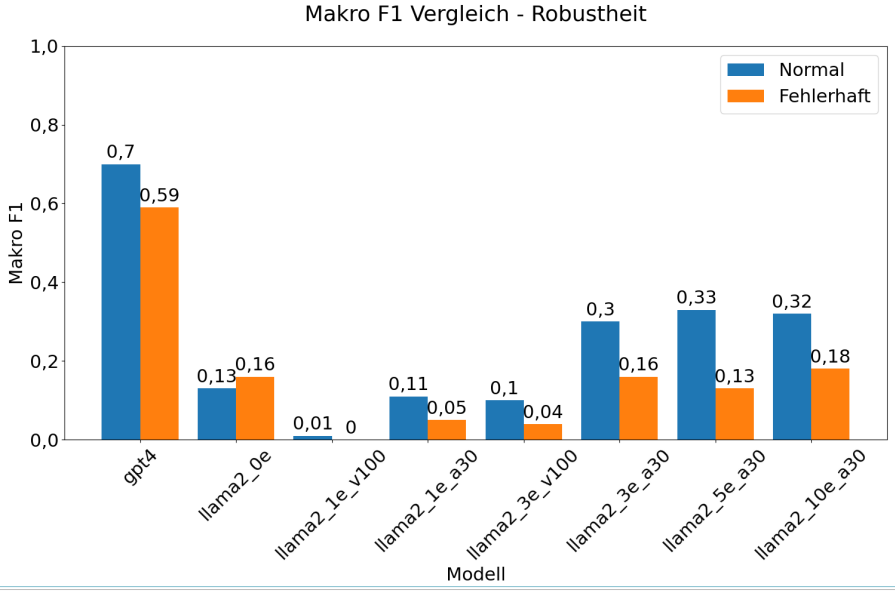
Kriterium Erklärbarkeit



Kriterium Fragenverständnis



Kriterium Robustheit



Ergebnisse - Vergleich

Modell	MakroF1	Erklärbarkeit	Fragenverständnis	Robustheit (Leistungsverlust)
GPT4	**0,7**	**97,3** %	**97,9** %	15,7 %
`llama2_0e`	0,13	48,1 %	64,2 %	−**0,23** %
`llama2_1e_v100`	0,01	0 %	2,1 %	100 %
`llama2_1e_a30`	0,11	50 %	54,7 %	54,5 %
`llama2_3e_v100`	0,1	47,4 %	40 %	60 %
`llama2_3e_a30`	0,3	**85,4** %	**72,6** %	46,7 %
`llama2_5e_a30`	**0,33**	82,5 %	60 %	60,6 %
`llama2_10e_a30`	0,32	81 %	55,8 %	**43,8** %

Diskussion

- Grenzen der Modelle
  - Kleine Modelle geben schlechtere Startbedingungen
  - Kleine Datensätze führen zu schneller Überanpassung
  - Kleiner Evaluierungsdatensatz keine hohe Signifikanz
- Probleme bei Kernfragen
  - Reproduzierung von falschem Wissen
  - Widersprüchliche Aussagen bei ähnlicher Eingabe
  - Verbesserung durch Überanpassung

Ausblick

Ansatz	Aufwand	Erfolgseinschätzung
Retrieval Augmented Generation	hoch	sehr hoch
Datensatzvergrößerung	niedrig	hoch
Modellvergrößerung	hoch	hoch
Human Reinforcement Learning	sehr hoch	hoch
Domänenspezifische Modelle	niedrig	mittel

Quellen

- Lewis, Patrick u. a. (2020). „Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks“. In: ArXiv abs/2005.11401. url: https://api.semanticscholar.org/CorpusID:218869575.
- Mao, Yuning, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han und Weizhu Chen (2020). „Generation-Augmented Retrieval for Open-Domain Question Answering“. In: Annual Meeting of the Association for Computational Linguistics. url: https://api.semanticscholar.org/CorpusID:221802772.
- Omar, Reham, Omij Mangukiya, Panos Kalnis und Essam Mansour (2023). ChatGPT versus Traditional Question Answering for Knowledge Graphs: Current Status and Future Directions Towards Knowledge Graph Chatbots. Version 1. arXiv: 2302.06466 [cs.CL].
- OpenAI (2023). „GPT-4 Technical Report“. In: arXiv: 2303.08774[cs.CL].
- Touvron, Hugo u. a. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv: 2302.13971 [cs.CL].
- Touvron, Hugo u. a. (2023b). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv: 2307.09288 [cs.CL].
- Winter, Alfred, Elske Ammenwerth, Reinhold Haux, Michael Marschollek, Bianca Steiner und Franziska Jahn (2023). Health Information Systems. 3. Aufl. Health Informatics. Springer Cham. isbn: 978-3-031-12310-8. doi: 10.1007/978-3-031-12310-8.

UNIVERSITÄT
LEIPZIG

Medizinische Fakultät

VIELEN DANK!

Paul Keller
Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE)
www.imise.uni-leipzig.de