GerMedIQ: A Resource for Simulated and Synthesized Anamnesis Interview Questions in German

Justin Hofenbitzer[1], Sebastian Schöning[2], Sebastian Belle[3], Jacqueline Lammert[1], Luise Modersohn[1], Martin Boeker[1], Diego Frassinelli[4]

[1]Technical University of Munich, [2]Fraunhofer IPA, [3]University of Heidelberg, [4]LMU Munich

justin.hofenbitzer@tum.de

The Problem: German Clinical Data is Sparse

- Strict **privacy regulations** in Germany and the EU for clinical data
- In the US, large datasets (e.g., MIMIC *(Johnson et al., 2016)*) contain real clinical texts
- No comparable corpora in German *(Hahn, 2025)*

The Solution: Data Augmentation

Synthetic data generation and **augmentation** of existing datasets with the help of LLMs *(Piedboef and Langlais, 2024)*

The German Medical Interview Questions Corpus (GerMedIQ)

Simulated Human Responses:

- 4,524 unique question-response pairs
- 116 German questions from **standardized medical anamnesis questionnaires** (University Medical Centre Mannheim)
- 39 laypersons answered each question with an *appropriate* response in German without disclosing any personally identifiable information

LLM-Augmented Synthetic Responses:

- 18 open-weight LLMs produced five independent responses per question in a stateless setup
- Zero-Shot Inference: No human responses were provided

	General Domain	Biomedical Domain	Total
Small *(< 3B)*	2	3	**5**
Medium	8	3	**11**
Large *(> 8B)*	2	0	**2**
Total	**12**	**6**	**18**

The GerMedIQ Corpus:

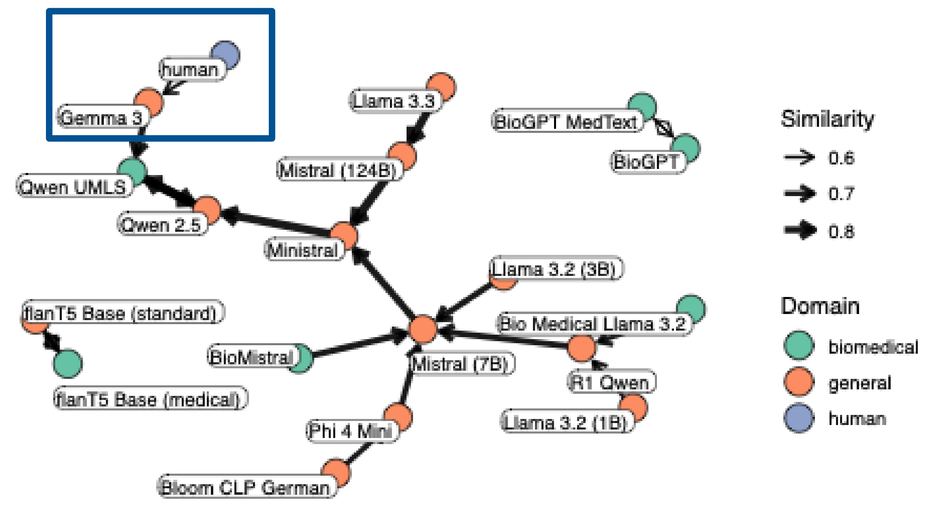
The Quality of the Dataset: A Take-Home Message

Structural Evaluation:

- Failing to follow instructions, 273 outputs from *BioGPT MedText* and *Gemma 3* were removed
- LLM responses were **longer** and **more complex** than human responses *(especially, small and medium-sized general models)*

Semantic Evaluation:

Gemma 3 was closest to the human responses, while small LLMs were the farthest away:



Acceptability Study:

- Humans (N = 4) and LLM judges (N = 15) rate the **acceptability** of each response given the corresponding question on a Likert scale between 1 (*completely unacceptable*) and 5 (*very acceptable*)
- LLMs and humans rated the responses from **medium and large LLMs better** than those from small models
- *Mistral (124B)* produced the **most acceptable** responses, **surpassing humans**

Leaderboard	Model	Count	Self-Vote	Human-Vote
Best	Mistral (124 B)	8/15	True	True
	Qwen 2.5	2/15	False	False
Worst	BioGPT	6/15	False	True
	BioGPT MedText	4/15	False	False

Universitäts Klinikum
Technical University of Munich
School of Medicine and Health
Institute for AI and Informatics in Medicine
TUM University Hospital

With funding from the:
Federal Ministry of Research, Technology and Space
Baden-Württemberg
MINISTERIUM FÜR WIRTSCHAFT, ARBEIT UND TOURISMUS

References:

- Udo Hahn. 2025. Clinical document corpora—real ones, translated and synthetic substitutes, and assorted domain proxies: a survey of diversity in corpus design, with focus on German text data. *JAMIA Open*.
- Alistair E. W. Johnson, et al.. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*.
- Frédéric Piedboeuf and Philippe Langlais. 2024. On Evaluation Protocols for Data Augmentation in a Limited Data Scenario. *arXiv preprint*.

ACL 2025 VIENNA
JULY 27 - AUGUST 1