GerMedIQ: A Resource for Simulated and Synthesized Anamnesis Interview Responses 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 frassinelli@cis.lmu.de

Abstract

Due to strict privacy regulations, text corpora in non-English clinical contexts are scarce. Consequently, synthetic data generation using Large Language Models (LLMs) emerges as a promising strategy to address this data gap. To evaluate the ability of LLMs in generating synthetic data, we applied them to our novel German Medical Interview Questions Corpus (GerMedIQ), which consists of 4,524 unique, simulated question-response pairs in German. We augmented our corpus by prompting 18 different LLMs to generate responses to the same questions. Structural and semantic evaluations of the generated responses revealed that large-sized language models produced responses comparable to those provided by humans. Additionally, an LLM-as-a-judge study, combined with a human baseline experiment assessing response acceptability, demonstrated that human raters preferred the responses generated by `Mistral (124B)` over those produced by humans. Nonetheless, our findings indicate that using LLMs for data augmentation in non-English clinical contexts requires caution.

1 Introduction

Textual medical data is crucial for developing and validating Natural Language Processing (NLP) applications within clinical contexts. While there are large, high-quality datasets available for English (e.g., MIMIC by Johnson et al. (2016)), accessible German clinical documentation typically remains sparse (Hahn, 2025). This is often due to stringent privacy constraints, restricted access to secure environments, or a lack of accessible corpora. While the creation of such shareable datasets should be viewed as the optimal solution, it is time-, labour-, and resource-intensive (Meineke et al., 2023; Lohr et al., 2024). A quicker and more lightweight alternative is data augmentation using Large Language Models (LLMs) (Piedboeuf and Langlais, 2024). However, the use of LLMs as robust *data generation engines* in the clinical domain remains largely underexplored, particularly regarding their capability to reliably simulate realistic clinical interactions between physicians and patients.

With this paper, we release the German Medical Interview Questions Corpus (GerMedIQ), a dataset consisting of 116 questions from standardized German anamnesis questionnaires and 39 simulated human responses each. Moreover, we explore the possibility of using LLMs in generating synthetic responses to those questions, specifically focusing on their ability to adopt the role of the patient.[1] The central question guiding our investigation is: Can LLMs effectively serve as synthetic data generators in the context of clinical anamnesis? Further, our experiments allow us to assess whether the same set of LLMs can also serve as judges.

2 Related Work

The following section provides an overview of existing medical interview datasets and dives deeper into the literature on synthetic data generation in the biomedical and clinical domains.

2.1 Medical Conversational Datasets

Researchers have collected real and simulated medical conversational datasets, mostly for training conversational artificial intelligence (AI) systems.

The largest real-world conversational dataset from the medical domain is MedDialog: Zeng et al. (2020) compiled a Chinese corpus with 3.4M doctor-patient interactions and an English corpus with 260K such conversations, covering numerous medical specialities. The researchers showed that models trained on the MedDialog dataset produced

[1]Throughout this paper, we differentiate between *simulated* and *synthetic* data: Both terms describe data that approximates real clinical data. We use the term *simulated* when the text was produced by humans, and *synthetic* whenever a machine generated it.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1064–1078

accurate medical conversations. Similar results are reported by Pieri et al. (2024) on models that were trained on BiMediX, a corpus combining 1.3M real and 200K synthetic English-Arabic clinical conversations. Xu et al. (2022) collected the RealMedDial dataset, consisting of 24K utterances from Chinese telemedical interviews, to train and improve medical dialogue systems. Saley et al. (2024) released a corpus of 22K English doctor-patient dialogues for medical history taking, and the dataset may serve task-oriented conversational AI systems. Another non-English corpus with Spanish counseling sessions includes 800 medical questions and about 400 expert reflections (Gunal et al., 2025). Gratch et al. (2014) collected the DAIC corpus with about 500 psychological English interviews for diagnosis support. The only medical interview corpus that includes German that we are aware of is DiK, which contains roughly 120 audio recordings with transcriptions of doctor-patient interactions in German, Portuguese, and Turkish as well as interpreted conversations to study interpretation in clinical multilingual scenarios (Bührig and Meyer, 2009).

In order to boost the automatic summarization abilities of LLMs as well as clinical note generation, Ben Abacha et al. (2023) collected a 1.7K corpus of simulated interactions between physicians and patients. Fareez et al. (2022) crafted a multimodal dataset consisting of 272 medical conversations derived from simulated cases focusing on respiratory diseases. Similarly, Papadopoulos Korfiatis et al. (2022) created a small, multimodal corpus for primary care consultations. Sanni et al. (2025) generated a dataset with medical and non-medical conversations in different African accents to enhance automatic speech recognition systems.

2.2 Synthetic Data Generation in the Biomedical Domain

The generation of synthetic data and the collection of simulated data have both evolved over the last years to overcome the shortage of clinical data caused by privacy constraints. Usually, data augmentation workflows are built upon existing data, where parts of datasets are paraphrased or back-translated by a model (Rentschler et al., 2022). Since the advancement of LLMs, researchers have been able to generate synthetic data completely independently from existing data sources, and Piedboeuf and Langlais (2024) showed that LLM-generated data increases model performance much better than paraphrasing or back-translations.

Typical reasons for the increasing interest in synthetic data generation are cost efficiency, scalability, control over the diversity and balance of data, and reduced privacy concerns, especially in healthcare (Liu et al., 2024; Nadas et al., 2025). This is underpinned by Hahn (2025), who states that besides domain proxies (e.g., guidelines) and translated real clinical datasets (e.g, in non-English contexts MIMIC-derived datasets), simulated or synthetic textual data are crucial for NLP applications in the clinical domain. Examples of existing German simulated text corpora are JSYNCC (Lohr et al., 2018) and GRASCCO (Modersohn et al., 2022).

A known disadvantage of LLM-generated data is their vulnerability to biases and hallucinations, potentially leading to counterfactual, unrealistic, or semantically implausible synthetic corpora (Yu et al., 2023; Hicks et al., 2024; Liu et al., 2024; Hahn, 2025; Nadas et al., 2025).

Synthetic data generation has been applied successfully in boosting LLMs' performance on arithmetics (Geva et al., 2020), information retrieval (Xiong et al., 2024), or named entity recognition (NER) (Lu et al., 2024). But also in the biomedical domain, data augmentation improved the performance of ICD-9 and ICD-10 code labeling (Kumichev et al., 2024; Sarkar et al., 2024) or other clinical NER tasks (Šuvalov et al., 2025); synthetic radiology reports helped to classify misdiagnosed fractures (Liu et al., 2025) and medical LLMs trained on synthetic text only even outperformed ones trained on real data (Peng et al., 2023).

3 Dataset: The GerMedIQ Corpus

We present the German Medical Interview Questions Corpus (GerMedIQ), consisting of 116 standardized anamnesis questions answered by 39 participants, resulting in 4,524 simulated unique German question-response pairs.[2] To the best of our knowledge, this is the first anamnesis interview question-response dataset for German.

3.1 The Corpus Collection

The interview questions were extracted from a mixture of standardized questionnaires and basic anamnesis questions used at the University Medical Centre Mannheim (UMM).

[2]The GerMedIQ Corpus and the LLM-augmented responses are available at Zenodo (https://www.doi.org/10.5281/zenodo.15774407) and GitHub (https://github.com/Jhofenbitzer/GerMedIQ-Corpus).

We selected the Barthel Index (Mahoney and Barthel, 1965), the EORTC Quality of Life Questionnaire (Aaronson et al., 1993), and the PainDETECT Questionnaire (Freynhagen et al., 2006), which are actively used in everyday clinical routines. The Barthel Index is designed to assess the functional abilities, e.g., mobility, to track changes in long-term patients. The EORTC Quality of Life Questionnaire is used to evaluate the physical, psychological, and social well-being of cancer patients. The PainDETECT Questionnaire screens neuropathic pain components in patients with chronic diseases. In addition, we compiled anamnesis questions from clinical routine interviews done at UMM covering a wide variety of topics like basic body characteristics, e.g., weight, or the medical history of a patient.[3] Some questions were slightly rephrased for consistency reasons.

Table 1 shows the distribution of questions across the full list of questionnaires. Due to privacy regulations, we could not collect responses from real patients and instead recruited laypeople without previous formal medical knowledge or known medical history. The rationale behind this decision is that no medical knowledge should be required to answer anamnesis questionnaires. In order to obtain realistic responses, the participants were instructed to give 'appropriate', i.e., grammatically well-formed and contextually reasonable responses without disclosing any personally identifiable information. Although no detailed patient profiles were provided, participants were encouraged to answer as plausibly as possible, drawing on their own understanding or interpretation of hypothetical clinical scenarios. All participants answered all questions online on MyMedax[4]. The survey took each participant roughly 40 minutes, and they received monetary compensation.

The GermMedIQ corpus contains three different question types: 12 Wh-questions (WhQ), 59 polar questions (PQ; yes/no-questions), and 39 questions that combine the two syntactic types (CQ). While PQ semantically denote a binary set of propositions (i.e., either confirming or rejecting the question), WhQ are known to have a significantly larger response space (e.g, cf. Hamblin, 1958, 1973; Karttunen, 1977; Groenendijk and Stokhof, 1984). Three sample questions per question type, together with potential responses, can be seen in (1) - (3).

Questionnaire	**N**
Baseline: Previous Medical History	19
Baseline: Anamnesis Assessment	16
Baseline: Basic (Subjective) History	16
EORTC QLQ 30	14
PainDetect Questionnaire	9
Barthel Index	8
Baseline: Patient Characteristics	7
Baseline: Patient Circumstances	7
Baseline: Immune System	6
Baseline: Senses	5
Baseline: Cardiovascular System	3
Baseline: Airways	2
Baseline: Existing Documents	2
Baseline: Teeth	1
Baseline: Upper Abdominal Organs	1
Total	116

Table 1: Distribution of questions per questionnaire.

(1) **Waren Sie kurzatmig?** (*Have you experienced shortness of breath?*)
   a. Ja (*Yes*)
   b. Nein, es gab keine Probleme (*No, there were no problems*)

(2) **Wie oft trinken Sie Alkohol pro Woche?** (*How often do you consume alcohol per week?*)
   a. Ich trinke zwei Bier (*I drink two beers*)
   b. Ich trinke nicht (*I don't drink*)

(3) **Üben Sie regelmäßig einen bestimmten Sport aus? Falls ja, bitte nennen Sie die Sportart** (*Do you exercise a specific sport regularly? If so, please specify which sport.*)
   a. Ich gehe regelmäßig schwimmen (*I go swimming regularly*)
   b. Ich spiele Tennis, dienstags im Verein (*I play tennis, every Tuesday with my club*)

3.2 Data Augmentation Process

We augmented the human-produced GerMedIQ corpus with machine-generated, synthetic responses from 18 open-weight LLMs without fine-tuning in a zero-shot approach. We selected a vanilla and, if existing, a biomedically fine-tuned variant of each LLM, ranging over different architectures and sizes. Table 2 summarizes the key

[3] Some of the baseline questionnaires are inspired by Kuhlmann et al. (2022) and the 'Deutscher Schmerzfragebogen Version 12/2024'.

[4] https://mymedax.de

characteristics of the models used.[5] Each model was instructed to respond to the upcoming anamnesis question as if it were a real patient. All models were exposed to the same prompt written in German, and we collected five independent responses from each model in a stateless setup.[6] Inference on an NVIDIA A40 48GB took overall ≈ 6.5 hours.

Model	Parameters	Domain	Size
flanT5 Base (standard)	250 M	general	S
flanT5 Base (medical)	250 M	biomedical	S
BioGPT	347 M	biomedical	S
BioGPT MedText	347 M	biomedical	S
Llama 3.2	1.0 B	general	S
Bio Medical Llama 3.2	1.0 B	biomedical	M
Llama 3.2	3.0 B	general	M
Llama 3.3	70.0 B	general	L
Phi 4 Mini	3.8 B	general	M
Gemma 3	4.0 B	general	M
Bloom CLP German	6.4 B	general	M
Qwen 2.5	7.0 B	general	M
Qwen UMLS	7.0 B	biomedical	M
R1 Qwen	8.0 B	general	M
Mistral	7.0 B	general	M
BioMistral	7.0 B	biomedical	M
Ministral	8.0 B	general	M
Mistral	124.0 B	general	L

Table 2: Overview of two encoder-decoder (flanT5) and 16 decoder-only models used for synthetic data generation.

4 Evaluation of synthetic data points

While it is straightforward to generate synthetic data with LLMs, the evaluation of the output has to be conducted carefully. To evaluate the quality of machine-generated responses and compare them with the human-generated ones, we performed two studies targeting structural and semantic properties of the output and one acceptability study.

4.1 Structural Evaluation

As a first approximation to the differences between human-produced and machine-generated responses to anamnesis interview questions, we measured the syntactic and grammatical properties of each type. In order to get realistic results, we decided to remove all model-internal tokens, e.g., end-of-sequence tokens, from the original strings of the synthetic LLM responses. If a response consisted exclusively of such tokens, we removed it from further analyses. In total, we filtered out 273 responses, 136 produced by `BioGPT MedText` and 137 by `Gemma 3` (cf. the last column in Table 3).

We used DoPaMeter (Lohr and Hahn, 2023) to retrieve the average number of tokens and characters, the type token ratio (TTR), as well as the average and maximum dependency distance from the responses. We aggregated the responses by *model domain*, *size*, *question type*, and all their interactions prior to computing the results.[7] While the token and character counts per response capture the average length of the given responses, TTR divides the number of distinct word forms by the total number of tokens and gives insights about the observed lexical diversity within the responses (Peirce, 1906). The average and maximum dependency distance measures the linear distance between all syntactic heads and their dependents and indicates how complex sentences are.

Table 3 shows that humans formulated shorter responses than models, regardless of their size, their domain, or the given question type. For example, human responses to PQ were about six tokens, while general-domain medium-sized LLMs produced answers of on average more than eleven, which is an increase of 83.3%. This trend is also reflected in the grammatical complexity, operationalized as the dependency distance: Human responses show lower average distances between syntactic heads and their dependents, indicating less complex sentence structures, compared to all groups of models. Moreover, responses to WhQ were on average about two tokens shorter and showed a lower average dependency distance than those to PQ or CQ for humans, medium, and large LLMs. The maximum dependency distance, i.e., the biggest distance between a token and its dominating head, does not show much variance for the answers given by humans (5.05-5.58), biomedical medium (6.59-7.81), and large LLMs (4.71-5.32). Small LLMs produce responses with higher complexity (general: 8.41-11.78, biomedical: 9.66-16.08), and medium-sized general-domain LLMs generated responses with very high maximum dependency distances (24.50-42.23). The evaluation of the lexical diversity in the responses did not reveal relevant differences.

[5] Model references are listed in Table 6 in Appendix A.1.
[6] Find the prompt in Figure 3 in Appendix A.2.
[7] We consider *small (S)* models having 1B or fewer parameters, *medium-sized (M)* models having more than 1B and up to 8B parameters, and *large (L)* models having more than 8B parameters (see column 'Size' in Table 2).

Domain	Q-Type	Size	N	Avg. Tokens	Avg. Characters	Avg. Dist.	Max. Dist.	TTR	Null
Humans	PQ	–	2301	6.38	32.34	1.40	5.58	0.14	–
	WhQ	–	819	4.62	24.76	0.99	5.05	0.30	–
	CQ	–	1404	6.65	35.45	1.45	5.15	0.19	–
General LLMs	PQ	S	590	8.93	46.21	1.90	11.22	0.23	–
		M	2609	11.08	58.27	2.14	42.08	0.11	46
		L	590	10.05	54.21	1.95	5.32	0.11	–
	WhQ	S	210	9.53	50.35	1.97	11.78	0.31	–
		M	921	10.16	52.88	1.99	42.23	0.17	24
		L	210	9.20	48.67	1.80	5.00	0.19	–
	CQ	S	360	9.76	51.80	1.99	8.41	0.25	–
		M	1553	10.53	58.11	2.07	24.50	0.15	67
		L	360	9.82	54.14	1.88	4.71	0.14	–
Biomedical LLMs	PQ	S	1108	8.67	44.20	1.73	16.08	0.25	72
		M	590	10.67	56.79	2.12	7.76	0.20	–
	WhQ	S	388	9.07	47.01	1.81	14.49	0.30	32
		M	210	9.52	49.58	1.96	6.59	0.29	–
	CQ	S	688	9.16	47.21	1.80	9.66	0.28	32
		M	360	9.76	52.60	2.03	7.81	0.24	–

Table 3: Overview of structural evaluation metrics: Amount of responses per evaluated group (N), Average amount of tokens and characters, average and maximum dependency distance, type token ratio (TTR) of given responses, and the number of detected null-responses (Null).

The structural evaluation showed `BioGPT MedText` and `Gemma 3` had trouble following the instructions, as 273 responses had to be removed from further analyses. Further, we saw that the remaining LLM responses were longer and, on average, more complex than the ones from the humans. Moreover, we showed that out of the most complex responses, those from humans were most consistent in having low complexity, together with medium-sized biomedical and large general models. This finding suggests that specifically small and medium-sized general models have produced oddly complex outlier responses.

4.2 Semantic Evaluation

In the second step of our investigation, we focused on the contextual relation between human and synthetic data via distributional semantics. Specifically, we looked into the diversity of responses per model, the similarity among models, and the closeness to human responses.

To analyze semantic similarity between responses, we used the `SentenceTransformers` library to compute sentence-level embeddings for each response (cf. Reimers and Gurevych, 2020).[8] We first computed *within-model similarity*, i.e., pairwise cosine similarity among all responses from the same model per question.[9] Second, we calculated *between-model similarity*, where we used cosine similarity between response centroids, i.e., the *average response*, to compare models with each other and with the human responses.

We fitted a series of linear mixed-effects regression models (LMER) on the within-model diversity using the `lme4` R-package (Bates et al., 2015). We compared all models with likelihood ratio tests to assess improvements in model fit. We began with a baseline intercept-only model including random intercepts for *question ID* and random slopes for *question type* by *model*, accounting for potential effects of single questions as well as question type preferences of the examined models. We then increased the complexity of the models by first adding *model domain*, *model size*, and *question type* as fixed effects. We then added two-way and, in the last model, three-way interactions between the predictors. The likelihood ratio comparison of the different models exhibited that the fixed-effects-only model provides the best fit (χ^2 = 13.9992, df = 6, p = .023).

The predictors of the chosen LMER revealed a significant positive effect of model size: Large (β = 0.282, p < .001) and medium models

[8] paraphrase-multilingual-MiniLM-L12-v2

[9] For simplicity reasons, we treat human responses as their own *model*, *domain*, and *size*.

(β = 0.153, p = .002) showed to have significantly higher within-model similarity scores, i.e., lower diversity in responses, than small models or humans (see Figure 1). Other fixed effects, including *question type* and *model domain*, did not reach statistical significance.

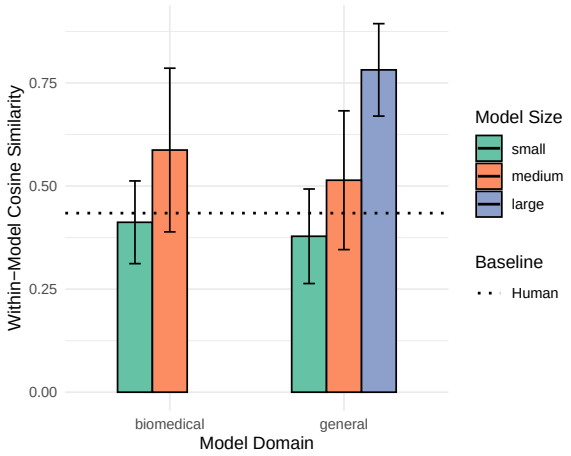

Figure 1: Within-model cosine similarity scores to account for diversity of responses of each model with standard deviation (for humans: $\pm$0.064). The figure divides the values by *model domain* and *model size*.

To account for between-model similarity, we calculated how far the centroid response of each LLM and the human responses deviated from those of all other models. Figure 2 displays a similarity graph where every model's centroid response is represented by a node. The arrows between the nodes reflect the between-model similarity and are directed to a centroid's most similar counterpart. The closest centroid response to the human centroid was produced by `Gemma 3` (cos = .63), a general-domain, medium-sized decoder-only LLM. Furthermore, we observe two similarity islands: Both `flanT5` models and the two small-sized GPT models, `BioGPT MedText` and `BioGPT`, produced very similar responses. Moreover, all models from the `Mistral` family are grouped together, and `Mistral (7B)`'s centroid was most similar to the largest number of other LLMs (N = 4).

While the human centroid was not among the top similar picks of any model, we further examined the distance between human and model centroids. We fitted a sequence of LMER using the same methodology as before. The structure of the model with the best fit (χ^2 = 16.2405, df = 5, p < .001) predicts the average centroid distance to the human centroid having *question type*, *model domain*, and *size* as non-interacting fixed-effects. Random effects were identical to the within-similarity LMER. The analysis of this model showed that, specifically, responses of large (β = -0.112, p < .001) and medium-sized LLMs (β = -0.075, p < .001) exhibited significantly lower distance to the human centroid than small models.

The semantic analysis of the human and machine responses revealed that small LLMs, as well as humans, produced more diverse responses than medium and large LLMs. By investigating the between-model distance, the human response centroid was not picked by any model as the most similar one, suggesting substantial semantic differences between human and LLM text. `Gemma 3` outperformed the other LLMs in getting closest to the human centroid, suggesting better ability to mimic humans. Two similarity islands and a cluster within the graph network indicate that more similar responses are produced within model families. On the other hand, `Mistral (7B)` was found to be most similar to most other models, where three out of four do not belong to the `Mistral` family. Lastly, the assessment of the distance of model centroids to the human's illustrated that small LLMs are the farthest away.

4.3 Acceptability Study

To assess the quality of human and machine responses, we conducted a human evaluation and an LLM-as-a-judge experiment (Zheng et al., 2023).

We asked four second-year medical students to rate the acceptability of a small subsample of the GerMedIQ corpus to ground the LLM judgments. All participants were native German speakers, and they passed the first medical state exam. Each question was extracted twice from the original corpus—once paired with a human response and once with a model-generated response—resulting in 232 unique question-response pairs. We further split the sample in half, each containing every question, making sure that 50% of the responses were generated by LLMs and 50% by humans. Two pseudo-randomized versions of each list were created, making sure that human responses and model responses were presented in alternating order, resulting in four experimental lists. Each human rater was presented with one of these lists and asked to judge the acceptability of each response on a Likert scale (Likert, 1932) from 1 (completely unacceptable) to 5 (very acceptable). Participants were instructed to assume acceptability if a response was *correct*, *natural*, and *contextually sound*.

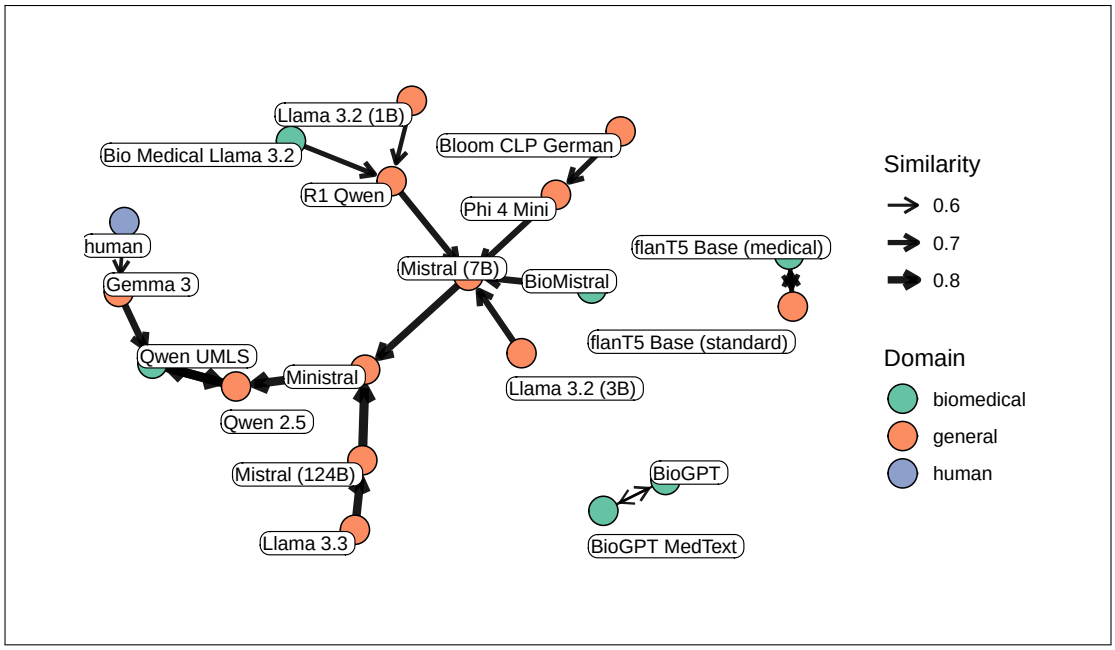

Figure 2: Semantic network graph displaying the highest centroid similarity for each model. The thickness of a connection indicates the similarity score.

This design ensures that each response was judged by two independent human evaluators. Each LLM, which was used as a data augmentor, was also instructed to judge the acceptability of every response given the respective question. The task was the same as for the humans and we constructed a unified English prompt describing the rating task carefully.[10] The models were instructed to respond with a single digit in the Likert-scale range only. We designed a zero-shot experiment with a stateless model setup to enhance comparability, and the overall runtime was $\approx$ 10 hours.

Substantial post-processing was necessary since many models did not comply with the instructions. We first removed every non-digit character from the judgments before we removed every number outside of the allowed range. This led to large exclusions of judgments (cf. Table 4), and we decided to exclude both `flanT5` models and `Gemma 3` from further analyses. We also removed all elements with fewer than two ratings, ending up with a total of 13,399 rated elements.

A post-hoc inter-rater agreement evaluation showed very low averaged pairwise Cohen's κ (Cohen, 1960) for both the human and the machine judgments, the latter being substantially lower (κ_{human} = .277; κ_{llm} = .055). After binarizing the ratings into *unacceptable* (ratings 1 to 3) and *acceptable* (ratings 4 and 5), we found moderate agreement for the humans and still low agreement for the LLMs (κ_{human} = .521; κ_{llm} = .144). Further analyses were conducted using the binary scores.

To examine the effects of model and judge characteristics on rating behavior, we employed a set of generalized linear mixed-effects regression models (gLMER) using `lme4`.

Model	Removed Outputs	N
Mistral (7B)	25.42%	3,406
Llama 3.2 (3B)	29.91%	4,008
Mistral (124B)	33.27%	4,458
Phi 4 Mini	35.58%	4,768
Qwen 2.5	44.30%	5,936
Qwen 2.5 UMLS	44.67%	5,986
Llama 3.3	49.47%	6,629
Minstral	56.97%	7,634
R1 Qwen	62.90%	8,428
Llama 3.2 (1B)	63.55%	8,515
BioMistral	66.07%	8,853
Bio Medical Llama 3.2	81.67%	10,943
Bloom CLP German	86.63%	11,608
BioGPT MedText	94.19%	12,620
BioGPT	97.10%	13,011
Gemma 3	99.97%	13,395
flanT5 Base (standard)	100.00%	13,399
flanT5 Base (medical)	100.00%	13,399

Table 4: Percentage and absolute count of removed judgments per model after post-processing due to instruction violations. The total number of judgments is 13,399.

[10] A comparison between the final prompt (cf. Figure 4 in Appendix A.2) and three alternatives—a direct German translation, a version requesting justification, and one requiring three ratings per criterion—revealed no notable differences in the judgments upon qualitative inspection of the results.

We replicated the procedure described in the semantic evaluation section and found our final model for the LLM judges (χ^2 = 2117.7, df = 8, p < .001) employing the binary rating score as the dependent variable modeled with a binomial distribution and a logit link. The fixed effects included *question type* and the interaction between *model and rater domain* as well as the interaction between *model and judge size*. Random intercepts were included for both the *question ID* and the *LLM judge* to account for question-specific and rater-specific variability. The final model for the human evaluators (χ^2 = 198.1347, df = 6, p < .001) included *question type*, *model domain*, and *size* as fixed effects without any interactions. The random-effects structure allowed random intercepts for *question ID* and *rater*, too.

The human gLMER revealed a significant negative main effect of *model domain*, i.e., responses from LLMs received lower ratings than human responses (e.g., for general LLMs: β = -5.487, OR = .004, p < .001). The LLM gLMER also shows a negative effect, indicating that general LLMs' answers were rated worse than humans' (β = -0.161, OR = 8.51, p < .001). A significant interaction between *model and judge domain* further clarifies that general-domain judges rated LLM responses better than biomedical judges, and thus LLMs received higher ratings than humans from general-domain judges (e.g., for general judges and general models: β = 0.299, OR = 1.35, p < .001). Moreover, both gLMER models revealed significant main effects of model size: large and medium models received significantly higher ratings compared to small models (e.g., for large models: β = 0.665, OR = 1.95, p < .001), also from human raters (e.g., for large models: β = 6.504, p < .001). Also, a significant negative effect of *judge size* was observed, indicating that large judges tended to give overall lower ratings than small-sized judges (β = -3.493, OR = .0304, p < .001). Similarly, the interaction between *model size* and *judge size* was highly significant in the LLM model: Human responses as well as those from medium and large LLMs received more favorable ratings from large and medium-sized judges than small LLMs (e.g., the interaction between large judges and large LLMs: β = 7.858, OR = 2588, p < .001). Question types were no significant predictor for the human ratings, while for LLMs, CQ were rated slightly lower than PQ (β = -0.095, OR = .909, p < .01).[11]

[11]For more details see Figures 5 and 6 in Appendix A.3.

We computed how often each judge rated each model being *acceptable* or *unaccpetable* and derived a leaderboard from the top-rated model per judge. Table 5 displays all models that were rated most and least appropriate more than once by transparently illustrating whether the respective model voted for itself and whether humans agreed with the top ranking. It can be seen that the responses from `Mistral (124B)` were perceived as most appropriate by most LLMs and the human raters. Also, the large Mistral model was the only one among the winners, which rated its own responses best. `Qwen 2.5` was rated most appropriate by two judges. The two `BioGPT` models were rated worst by 10 out of 15 LLMs, plus the humans, indicating low performance. It is surprising, though, that neither the LLM judges nor the human evaluators rated the human responses as most acceptable.

	Model	Count	Self-vote	Human Vote
Best	Mistral (124B)	8/15	T	T
	Qwen 2.5	2/15	F	F
Worst	BioGPT	6/15	F	T
	BioGPT MedText	4/15	F	F

Table 5: Leaderboard of the rated models: Count of best and worst rated models by all LLM and human judges, including self-votes.

This study showcased once more that LLMs do not always follow the given instructions, which led to the exclusion of three models in the LLM-as-a-judge study. To enhance agreement within both human raters and LLM judges, we binarized the rating scores. The analyses demonstrated different preferences: While humans and biomedical models classified human responses as more appropriate compared to LLM responses, general-domain models held the inverse point of view. Correspondingly, question type was no significant factor for humans, while LLM judges rated responses to CQ worse than to PQ or WhQ. LLMs and humans agreed that large and medium-sized LLMs produced more appropriate responses than small models. Also, large judges were shown to rate all responses more conservatively than small-sized judges. In addition, `Mistral (124B)` was rated most appropriate by the majority of LLM judges and, surprisingly, also by the human raters, while the two `BioGPT` models produced the most inappropriate responses, according to all judgments.

5 General Discussion

The driving question behind the three evaluation studies was to identify whether open-weight LLMs serve as reliable synthetic data generators. Before even evaluating the synthetic responses, we found that a small portion of the given responses by `BioGPT MedText` and `Gemma 3` had to be removed from further analyses. Even worse was the situation with the LLM-as-a-judge study, where no LLM fully complied with the instructions given, and both `flanT5`'s and `Gemma 3` had to be excluded. We assume that one reason for this finding is the lack of model-specific prompts. Recent research found that even state-of-the-art models show significant vulnerability of LLMs when used as judges (Maloyan et al., 2025).

Furthermore, the structural, semantic, and acceptability evaluations indicated a clear pattern: Especially large LLMs, but mostly also medium-sized ones, perform at least on par with humans. While humans distinctly produced shorter and less complex responses than all LLMs, medium-sized biomedical, and large LLMs, produced equally readable sentences as humans. The semantic evaluation further showed that medium and large LLMs synthesized responses significantly closer to the human answers than small LLMs, `Gemma 3` outperforming all other models. Finally, LLM judges and human raters agreed that small models' answers were significantly less acceptable. Moreover, the `BioGPT` models' responses were rated unacceptable most often, suggesting a larger quality gap.

Most surprisingly, though, were not human responses, but those from `Mistral (124B)`, the largest, general-domain model in our setup, rated to produce the most acceptable responses over all questions contained in our dataset. While, in general, humans rated human responses better than LLM responses, they agreed with the LLM judges that `Mistral (124B)` delivered the best responses to the questions. This finding supports recent investigations showing that LLMs are capable of outperforming humans across different domains and tasks (e.g., cf. Taloni et al., 2023; Marco et al., 2025; Salvi et al., 2025).

Altogether, the experiments showed that the use of LLMs for data augmentation in the context of German clinical language is possible once the right LLM has been identified. In our setup, `Gemma 3` was semantically closest to the human responses, and `Mistral (124B)` was rated to produce the most acceptable texts. We nevertheless think that a life-cycle for synthetic textual data or a human-in-the-loop approach might be important to consider before further processing LLM-augmented data, especially given the instruction compliance issue we found (cf. Liu et al., 2024; Long et al., 2024). In addition, we clarified that a fairly large and diverse set of LLMs can effectively be used in an LLM-as-a-judge setup, as their ratings largely agree with those from human raters. We did not identify biases when models judge their own responses.

6 Conclusion

We release a novel simulated medical anamnesis interview question dataset along with the synthetically generated responses by the LLMs, unique in the German clinical NLP environment. The dataset has the potential to improve conversational AI in health care and to give insights into the answering behaviour of both humans and LLMs.

Moreover, we could show that especially small LLMs should only be leveraged carefully as synthetic data generators in the German clinical context. Medium and large LLMs showed similar performance to humans across evaluations, with `Mistral (124B)` even outperforming humans in the rating study.

Future research should investigate further whether LLMs behave similarly in other non-English contexts, perhaps including closed-weight models and different architectures. In addition, prompt-tuning might be a valuable extension for both the data augmentation process and the LLM-as-a-judge experiment.

Acknowledgements

We thank Miriam Butt, Elena Schweizer, Lena Bitzer, Steffen Frenzel, Claudio Benzoni, Suteera Seeha, Sihan Wu, Leen Houri, Peter Pallaoro, Viktoria Hartmann, Lena Maria Zolda, Felicitas de la Cruz-Rothenfusser, and Elena Thias, who gave helpful feedback, and helped us with the recruitment process. We are very thankful that the Department of Linguistics at the University of Konstanz and the Digital Healthcare and Process Intelligence Group at Fraunhofer IPA allowed us to use their facilities for the corpus collection process. Furthermore, we would like to thank the Institute of AI and Informatics in Medicine at the Technical University of Munich for allowing us to use their computational resources for our study.

This research was supported by the German Federal Ministry of Research, Technology, and Space under the grant number 01ZZ2314A, and by the Ministry of Economic Affairs, Labor, and Tourism of the State of Baden-Württemberg under the grant number WM35-42-76/39/3.

Limitations

Due to privacy constraints in healthcare, our GerMedIQ corpus consists of simulated responses only. Therefore, evaluations based on the collected responses have to be done carefully, and comparing them to real patients' answers might increase their value further.

Both our data augmentation approach and the LLM-as-a-judge study leveraged a similar prompt for all models. Identifying optimal prompts for individual models, e.g., via soft-prompting or prompt-tuning, might lead to more accurate results. Also, we obtained only one round of judgments from each model. To estimate variability, multiple judgment rounds could be beneficial and represent more balanced ratings. Moreover, all LLM judges rated every response, including their own. While the highest rated model (`Mistral (124B)`) voted for itself, it would have won even without that vote. Nonetheless, we can not fully exclude model biases, which should be accounted for in follow-up experiments.

For the human evaluation, we only recruited four participants. Therefore, the presented results might be substantially influenced by subjective perceptions of individual raters. A replication of our study with a larger sample size would yield more reliable results.

Ethics Statement

We do not see any significant ethical issues related to this work. All our experiments involving human participants were conducted voluntarily with fair compensation. Participants in the corpus collection process received 5 Euros each, and the human raters were, at the time of the study, employed as research assistants. All participants were informed on how the data would be used, and we did not collect any information that could link the participants to the data. The corpus collection process was in line with the ethical regulations of the University of Konstanz (IRB number 05/2021). All our experiments were conducted with open-source libraries, which received due citations.

Use of AI Assistants

The authors acknowledge the use of ChatGPT for grammatical and stylistic enhancements of the final manuscripts, and for providing assistance with identifying the final prompts.

References

Neil K. Aaronson, Sam Ahmedzai, Bengt Bergman, Monika Bullinger, Ann Cull, Nicole J. Duez, Antonio Filiberti, Henning Flechtner, Stewart B. Fleishman, Johanna C. J. M. de Haes, Stein Kaasa, Marianne Klee, David Osoba, Darius Razavi, Peter B. Rofe, Simon Schraub, Kommer Sneeuw, Marianne Sullivan, and Fumikazu Takeda. 1993. The European Organization for Research and Treatment of Cancer QLQ-C30: A Quality-of-Life Instrument for Use in International Clinical Trials in Oncology. 85(5):365–376.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Kristin Bührig and Bernd Meyer. 2009. Dolmetschen im Krankenhaus (DiK).

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, and 12 others. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Preprint*, arXiv:2501.12948.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, Thomas Lo, and Christopher W. Smith. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1):313.

Rainer Freynhagen, Ralf Baron, Ulrich Gockel, and Thomas R. Tölle. 2006. Pain DETECT: a new screening questionnaire to identify neuropathic components in patients with back pain. 22(10):1911–1920.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting Numerical Reasoning Skills into Language Models. arXiv:2004.04487.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The Distress Analysis Interview Corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jeroen Groenendijk and Martin Stokhof. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis.

Aylin Ece Gunal, Bowen Yi, John D. Piette, Rada Mihalcea, and Veronica Perez-Rosas. 2025. Examining Spanish Counseling with MIDAS: a Motivational Interviewing Dataset in Spanish. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 866–872, Albuquerque, New Mexico. Association for Computational Linguistics.

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*, 8(3):ooaf024.

Charles L. Hamblin. 1958. Questions. 36:159–68.

Charles L. Hamblin. 1973. Questions in Montague English. 10(1):41–53.

Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. ChatGPT is bullshit. 26(2).

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Lauri Karttunen. 1977. Syntax and semantics of questions. 1(1):3–44.

Louise Kuhlmann, Keith Teo, Søren Schou Olesen, Anna Edwards Phillips, Mahya Faghih, Natalie Tuck, Elham Afghani, Vikesh K. Singh, Dhiraj Yadav, John A. Windsor, and Asbjørn Mohr Drewes. 2022. Development of the Comprehensive Pain Assessment Tool Short Form for Chronic Pancreatitis: Validity and Reliability Testing. *Clinical Gastroenterology and Hepatology*, 20(4):e770–e783.

Gleb Kumichev, Pavel Blinov, Yulia Kuzkina, Vasily Goncharov, Galina Zubkova, Nikolai Zenovkin, Aleksei Goncharov, and Andrey Savchenko. 2024. MedSyn: LLM-Based Synthetic Medical Text Generation Framework. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, pages 215–230, Cham. Springer Nature Switzerland.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. *Preprint*, arXiv:2402.10373.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140:55–55.

Jinghui Liu, Bevan Koopman, Nathan J. Brown, Kevin Chu, and Anthony Nguyen. 2025. Generating synthetic clinical text with local large language models to identify misdiagnosed limb fractures in radiology reports. *Artificial Intelligence in Medicine*, 159:103027.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best Practices and Lessons Learned on Synthetic Data for Language Models. *Preprint*, arXiv:2404.07503.

Christina Lohr, Sven Buechel, and Udo Hahn. 2018. Sharing Copies of Synthetic Clinical Corpora without Physical Distribution — A Case Study to Get Around IPRs and Privacy Constraints Featuring the German JSYNCC Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Christina Lohr and Udo Hahn. 2023. DOPA METER — A Tool Suite for Metrical Document Profiling and Aggregation.

Christina Lohr, Franz Matthies, Jakob Faller, Luise Modersohn, Andrea Riedel, Udo Hahn, Rebekka Kiser, Martin Boeker, and Frank Meineke. 2024. *De-Identifying GRASCCO - A Pilot Study for the De-Identification of the German Medical Text Project (GeMTeX) Corpus*, volume 317, pages 171–179. IOS Press.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. *Preprint*, arxiv:2406.15126.

Qiuhao Lu, Rui Li, Andrew Wen, Jinlian Wang, Liwei Wang, and Hongfang Liu. 2024. Large Language Models Struggle in Token-Level Clinical Named Entity Recognition. *arXiv preprint*. ArXiv:2407.00731.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for

biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6). Bbac409.

Florence I. Mahoney and Dorothea W. Barthel. 1965. Functional evaluation: the Barthel Index: a simple index of independence useful in scoring improvement in the rehabilitation of the chronically ill.

Narek Maloyan, Bislan Ashinov, and Dmitry Namiot. 2025. Investigating the Vulnerability of LLM-as-a-Judge Architectures to Prompt-Injection Attacks. *arXiv preprint*. ArXiv:2505.13348 [cs] version: 1.

Guillermo Marco, Luz Rello, and Julio Gonzalo. 2025. Small Language Models can Outperform Humans in Short Creative Writing: A Study Comparing SLMs with Humans and LLMs. *arXiv preprint*. ArXiv:2409.11547 [cs].

Frank Meineke, Luise Modersohn, Markus Loeffler, and Martin Boeker. 2023. Announcement of the German Medical Text Corpus Project (GeMTeX).

Luise Modersohn, Stefan Schulz, Christina Lohr, and Udo Hahn. 2022. GRASCCO - The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus. *Studies in Health Technology and Informatics*, 296:66–72.

Mihai Nadas, Laura Diosan, and Andreea Tomescu. 2025. Synthetic Data Generation Using Large Language Models: Advances in Text and Code. *arXiv preprint*. ArXiv:2503.14023 [cs].

Malte Ostendorff and Georg Rehm. 2023. Efficient Language Model Training through Cross-Lingual and Progressive Transfer Learning. *arXiv preprint*.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: A Dataset Of Primary Care Mock Consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Charles Santiago Sanders Peirce. 1906. *Prolegomena to an Apology for Pragmaticism*. The Monist.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smith, Nima PourNejatian, Anthony B. Costa, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Gloria Lipori, Duane A. Mitchell, Naykky S. Ospina, Mustafa M. Ahmed, William R. Hogan, Elizabeth A. Shenkman, Yi Guo, Jiang Bian, and Yonghui Wu. 2023. A study of generative large language model for medical research and healthcare. *npj Digital Medicine*, 6(1):1–10.

Frédéric Piedboeuf and Philippe Langlais. 2024. On Evaluation Protocols for Data Augmentation in a Limited Data Scenario. *arXiv preprint*. ArXiv:2402.14895 [cs].

Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. BiMediX: Bilingual Medical Mixture of Experts LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16984–17002, Miami, Florida, USA. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A Party of Foundation Models.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sophie Rentschler, Martin Riedl, Christian Stab, and Martin Rückert. 2022. Data Augmentation for Intent Classification of German Conversational Agents in the Finance Domain. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 1–7. KONVENS 2022 Organizers.

Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, and Mausam . 2024. MediTOD: An English Dialogue Dataset for Medical History Taking with Comprehensive Annotations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16843–16877, Miami, Florida, USA. Association for Computational Linguistics.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2025. On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, pages 1–9.

Mardhiyah Sanni, Tassallah Abdullahi, Devendra Deepak Kayande, Emmanuel Ayodele, Naome A Etori, Michael Samwel Mollel, Moshood O. Yekini, Chibuzor Okocha, Lukman Enegi Ismaila, Folafunmi Omofoye, Boluwatife A. Adewale, and Tobi Olatunji. 2025. Afrispeech-Dialog: A Benchmark Dataset for Spontaneous English Conversations in Healthcare and Beyond. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8399–8417, Albuquerque, New Mexico. Association for Computational Linguistics.

Atiquer Rahman Sarkar, Yao-Shun Chuang, Noman Mohammed, and Xiaoqian Jiang. 2024. De-identification is not enough: a comparison between de-identified and synthetic clinical notes. *Scientific Reports*, 14(1):29669.

Hendrik Šuvalov, Mihkel Lepson, Veronika Kukk, Maria Malk, Neeme Ilves, Hele-Andra Kuulmets, and Raivo Kolde. 2025. Using Synthetic Health Care Data to Leverage Large Language Models for Named Entity Recognition: Development and Validation Study. *J Med Internet Res*, 27:e66279.

Andrea Taloni, Massimiliano Borselli, Valentina Scarsi, Costanza Rossi, Giulia Coco, Vincenzo Scorcia, and Giuseppe Giannaccare. 2023. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. *Scientific Reports*, 13(1):18562.

Zheyang Xiong, Vasilis Papageorgiou, Kangwook Lee, and Dimitris Papailiopoulos. 2024. From Artificial Needles to Real Haystacks: Improving Retrieval Capabilities in LLMs by Finetuning on Synthetic Data. *Preprint*, arxiv:2406.19292.

Bo Xu, Hongtong Zhang, Jian Wang, Xiaokun Zhang, Dezhi Hao, Linlin Zong, Hongfei Lin, and Fenglong Ma. 2022. RealMedDial: A Real Telemedical Dialogue Dataset Collected from Online Chinese Short-Video Clips. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3342–3352, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: a tale of diversity and bias. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 55734–55784, Red Hook, NY, USA. Curran Associates Inc.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale Medical Dialogue Datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Appendix A.1

As an extension to Table 2, Table 6 lists all available research papers and Huggingface repositories of the LLMs described in this paper.

Appendix A.2

Figures 3[12] and 4 illustrate the final prompts used for the data augmentation and the LLM-as-a-judge experiment, respectively. While we did not perform quantitative prompt tuning, we always tried multiple prompts and subjectively assessed the quality of model outputs. Different prompts included a German and an English version. In the judgment study, we experimented with prompts asking the models to justify their response, or giving independent ratings for each appropriateness criterion mentioned in the prompt. For the data augmentation process, the German prompt turned out to be best, while for the LLM-as-a-judge experiment, the English, single-digit prompt worked best.

Data Augmentation Prompt:
"Du erhältst gleich eine Interviewfrage aus einem standardisierten medizinischen Anamnesefragebogen auf deutsch. Bitte beantworte die Frage auf deutsch, so als wenn du ein realer Patient in der Routineversorgung wärst.
Frage: {question}
Antwort:"

Figure 3: Plain-text LLM prompt for data augmentation.

LLM-as-a-judge Prompt:
"You are an expert in medical interviews and your task is to evaluate the quality of a given response to a medical questionnaire question, both written in German. Your rating should consider the appropriateness of a response. A response is considered appropriate if it answers the question properly, it is natural, coherent and contextually suitable. Rate each response on a scale from 1 (not appropriate) to 5 (very appropriate). Please, respond only with a number and do not justify your rating.
Question: {question}
Answer: {answer}
judgment:"

Figure 4: Plain-text LLM prompt for the LLM-as-a-judge study.

[12]English Translation: *You will immediately receive an interview question from a standardized anamnesis questionnaire in German. Please answer the question in German as if you were a real patient in routine care.*
Question: {question}.
Response:.

Appendix A.3

Figure 5 visualizes the average ratings of the human raters. Human responses were rated drastically higher, and small model responses much lower than by the LLM judges (cf. Figure 6), but the overall trend is similar: Large general-domain LLMs were rated best, and even higher than the human responses.

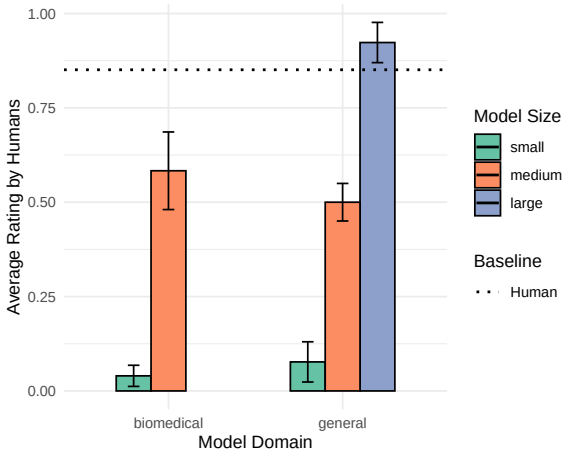

Figure 5: Average binary rating by human raters divided by model size and domain with standard error. The human standard error is ± 0.024

Figure 6 displays the mean ratings given by the LLM judges grouped by size and domain of judges as well as models. The figure visually represents the findings described in section 4.3 and showcases, for example, that large LLM judges preferred the responses of large models, even more than biomedical judges. Moreover, it is visible that medium LLMs were always rated higher than small LLMs, and large LLMs than medium-sized models.

Model	Huggingface Repository	Reference
flanT5 Base (standard)	google/flan-t5-base	Chung et al. (2022)
flanT5 Base (medical)	QuyenAnhDE/flant5base-medical	-
BioGPT	microsoft/biogpt	Luo et al. (2022)
BioGPT MedText	AventIQ-AI/BioGPT-MedText	-
Llama 3.2 (1B)	meta-llama/Llama-3.2-1B	-
Bio Medical Llama	ContactDoctor/Bio-Medical-Llama-3-2-1B-CoT-012025	-
Llama 3.2 (3B)	meta-llama/Llama-3.2-3B-Instruct	-
Llama 3.3	meta-llama/Llama-3.3-70B-Instruct	-
Phi 4 Mini	microsoft/Phi-4mini-instruct	-
Gemma 3	google/gemma-3-4b-it	-
Bloom CLP German	malteos/bloom-6b4-clp-german	Ostendorff and Rehm (2023)
Qwen 2.5	Qwen/Qwen2.5-VL-7B-Instruct	Yang et al. (2024); Qwen Team (2024)
Qwen UMLS	prithivMLmods/Qwen-UMLS-7B-Instruct	-
R1 Qwen	deepseek-ai/DeepSeek-R1-0528-Qwen3-8B	DeepSeek-AI (2025)
Mistral (7B)	mistralai/Mistral-7B-Instruct-v0.1	-
BioMistral	BioMistral/BioMistral-7B	Labrak et al. (2024)
Ministral	mistralai/Ministral-8B-Instruct-2410	-
Mistral (124B)	mistralai/Mistral-Large-Instruct-2411	-

Table 6: LLMs and their corresponding sources.

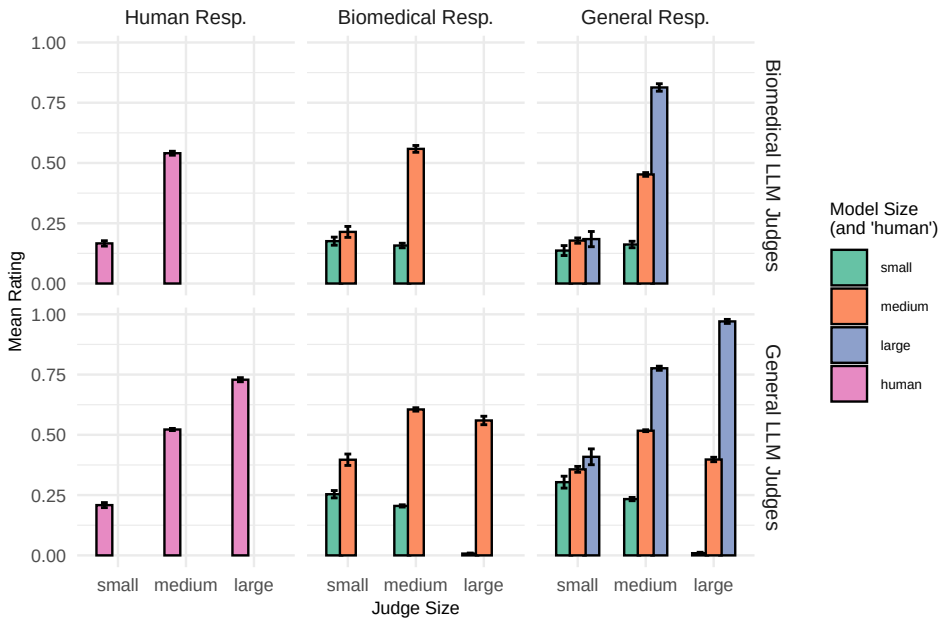

Figure 6: Average binary rating by LLM judges divided by judge and model size as well as judge and model domain with standard error.