# GerMedIQ: At the Gap Between Human and Synthetic Clinical Text

Anonymous ACL submission

## Abstract

Text corpora in non-English clinical contexts are sparse, where synthetic data generation with Large Language Models (LLMs) appears as a promising strategy to overcome this data gap. In order to test the quality of LLMs in generating synthetic data, we applied them to our novel German Medical Interview Questions Corpus (GerMedIQ), consisting of 4,524 unique question-response pairs in German. We augmented our corpus by asking a cohort of models to produce suitable responses to the same questions. Structural and semantic evaluations of the synthetic responses revealed that while augmented responses may meet the grammatical requirements, most models were not able to produce semantically comparable responses to humans. Also, an LLM-as-a-judge experiment showcased that human responses were consistently rated more appropriate than synthetic ones. We find that data augmentation with LLMs in non-English and clinical domain contexts has to be performed carefully.

## 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 the English language (e.g., MIMIC by Johnson et al. (2016)), accessible German clinical documentation typically remains sparse (Hahn, 2024). This is often due to stringent privacy constraints, restricted access to secure environments, or a lack of accessible datasets or corpora. While the creation of such shareable datasets should be viewed as the optimal solution, it is time-, labour-, and resource-intensive (cf. Meineke et al., 2023; Lohr et al., 2024). A quicker and more lightweight alternative is data augmentation using Large Language Models (LLMs) (cf. Piedboeuf and Langlais, 2024). However, the feasibility of using 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 real-world questions from standardized German anamnesis questionnaires and 39 simulated responses, each by humans. Moreover, we explore the suitability of small to medium-sized LLMs in generating synthetic responses to those questions, specifically focusing on their ability to adopt the role of the patient. 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 the ability of LLMs to simulate patient behaviour.

## 2 Related Work

The following section dives deeper into the existing literature on synthetic data generation in the biomedical and clinical domain. Furthermore, an overivew about existing medical interview datasets will be given.

### 2.1 Synthetic Data Generation in the Biomedical Domain

The generation of synthetic data has evolved over the last years to overcome shortages of clinical data due to privacy constraints. Common use cases are to augment smaller datasets with synthetic data points in order to increase their size or diversity, or to become more independent of time-consuming data collection processes. Usually, data augmentation workflows are built upon existing data, where parts of datasets are paraphrased or back-translated by a model (cf. Rentschler et al., 2022). Since the advancement of LLMs in recent years, researchers have been able to generate synthetic data com-

pletely independently from existing data sources, and Piedboeuf and Langlais (2024) showed that LLM-generated synthetic data tends to increase model performance much better than traditional approaches.

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 (2024), who states that besides domain proxies (e.g., guidelines) and translated datasets (e.g, in non-English contexts MIMIC-derived datasets), synthetic textual data are crucial for NLP applications in the clinical domain. Examples of existing German synthetic text corpora are JSYNCC (Lohr et al., 2018) and GRASCCO (Modersohn et al., 2022).

A known disadvantage of data synthesized by LLMs is the fact that those models are reportedly vulnerable to biases or hallucinations, potentially leading to counterfactual, unrealistic, or semantically implausible synthetic corpora (Yu et al., 2023; Liu et al., 2024; Hicks et al., 2024; Hahn, 2024; 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 (Sarkar et al., 2024; Kumichev 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).

### 2.2 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 accurate medical conversations. Similar results are reported by Pieri et al. (2024) on models that were trained on BiMediX, their 1.3M corpus of English-Arabic clinical conversations. Moreover, Xu et al. (2022) collected the RelMedDial dataset consisting of 24K utterances from Chinese telemedical interviews in order to train or improve medical dialogue systems. Saley et al. (2024) captured a 22K corpus of 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 included 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. DiK was collected to study the multilingual interpretation in the clinical context (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.

## 3 Dataset: The GerMedIQ Corpus

To the best of our knowledge, there is no available German anamnesis question dataset for research. Therefore, we present the German Medical Interview Questions Corpus (GerMedIQ), consisting of 116 standardized anamnesis questions answered by 39 participants, resulting in 4,524 unique German question-response pairs.

### 3.1 The Corpus Collection

The interview questions were extracted from a mixture of standardized questionnaires and basic anamnesis questions used at the University Hospital REDACTED. We took questions from 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 routine, especially with cancer patients. In addition, we compiled anamnesis questions from clinical routine interviews covering a wide variety of topics like basic body characteristics (e.g., weight or height) or the medical history of a patient. 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 were not able to collect responses from real patients and decided to focus on a cohort of laypeople without previous formal medical knowledge, and we have no information about their medical history. A rationale behind this decision is that no medical knowledge should be required to answer anamnesis questionnaires properly. In order to obtain realistic responses, the participants were explicitly instructed to give 'appropriate', i.e., grammatically well-formed and contextually reasonable responses to each question 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. The survey was conducted online on the platform MyMedax[1] and took each participant roughly 40 minutes. Each participant received monetary compensation to increase their level of motivation.

| **Questionnaire** | **N** |
|---|---|
| Baseline: Medical History | 19 |
| Baseline: Anamnesis Assessment | 16 |
| Baseline: 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.

The corpus[2] contains three different question types: 12 Wh-questions (WhQ), 59 polar questions (PQ) (or yes/no-questions), and 39 questions that combine the two syntactic question types (CQ). While PQ semantically denote a binary set of propositions (i.e., either confirming or rejecting the question), WhQ are known to have significantly larger response space (e.g, cf. Hamblin, 1958, 1973; Groenendijk and Stokhof, 1984; Karttunen, 1977). As CQ contain both question types, they are expected to behave similarly to WhQ. Thus, we hypothesize that WhQ and CQ evoke more diverse and longer responses compared to binary PQ. Three sample questions per question type, together with potential responses, can be seen in (1) - (3).

(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 15 LLMs in order to assess their quality without finetuning 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. As a use case, we focused on small to medium-sized, open-weight LLMs to guarantee broader accessibility. Table 2 displays the key characteristics of the mod-

[1] https://mymedax.de

[2] We will release the corpus upon acceptance.

els used.[3]

Each model was instructed to respond to the upcoming standardized medical anamnesis question as if it were a real patient being asked that question. All models were exposed to the same prompt written in German, and we collected five independent responses from each model. The LLMs were inferenced on a NVIDIA A40 48GB.

## 4 Evaluation of synthetic data points

While it is straightforward to generate synthetic data points with LLMs, the evaluation of the data has to be conducted carefully. To evaluate the quality of machine-generated responses, we conduct three complementary studies—linguistic-structural, semantic, and subjective rating—and compare the findings to those from human-generated responses.

### 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.

#### 4.1.1 Methods

To evaluate the structural features of synthetic and human responses, we first replaced those consisting solely of model-internal tokens with the placeholder `<EDIT-NO-RESPONSE>` before we used spaCy (Honnibal et al., 2020) to tokenize the responses. Moreover, we computed traditional corpus linguistic metrics, like response lengths, and token n-grams.

#### 4.1.2 Results

Aggregated over all questions, human responses have an average length of 6.64 tokens (min = 1, max = 60) or 34.61 characters (min = 1, max = 344). The longest responses received on average CQ with a mean length of 7.34 tokens or 39.22 characters (min = 1tk/1chr, max = 60tk/344chr). Human responses to WhQ (avg = 4.90tk/26.36chr, min = 1tk/1chr, max = 44tk/214chr) were in average shorter than those to PQ (avg = 6.83tk/34.73chr, min = 1tk/1chr, max = 52tk/301chr). The most frequent uni-, bi-, and trigrams in the human responses are *ich* ('I'), *ich habe* ('I have'), and *ja ich habe* ('yes I have').

The average synthetic responses are longer than the human responses (avg = 27.18tk/148.51chr, min = 1tk/1chr, max = 533tk/563chr). Responses by biomedical models are slightly shorter (avg = 24.13tk/131.59chr, min = 1tk/1chr, max = 64tk/270chr) compared to those generated by general LLMs (avg = 29.21tk/159.79chr, min = 1tk/1chr, max = 533tk/563chr). The model with the longest responses over all categories is `Ministral-8B-Instruct-2410` (avg = 36.43tk/204.31chr, min = 4tk/13chr, max = 103tk/371chr), but even `BioGPT-MedText`, the model with the shortest responses over all questions, has on average longer responses than humans (avg = 7.80tk/37.46chr, min = 1tk/1chr, max = 48tk/233chr). The most frequent n-grams within the synthetic responses are identical to the human n-grams.[4].

#### 4.1.3 Interim Discussion

The structural evaluation showed that LLMs tend to produce more lengthy responses to anamnesis interview questions than humans, with general LLMs being slightly longer than biomedical ones. Interestingly, the shortest human responses were found with WhQ, contradicting the initial hypothesis based on the idea that WhQ would have longer responses due to their increased semantic response space. Despite differences in the response length, the most frequent n-grams of the LLMs correspond to the ones found in the human responses, suggesting that LLMs correctly identified relevant keywords for this task.

### 4.2 Semantic Evaluation

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

#### 4.2.1 Methods

To analyze semantic similarity between responses, we used the SentenceTransformers library with the `paraphrase-multilingual-MiniLM-L12-v2` model to compute sentence-level embeddings for each response (cf. Reimers and Gurevych, 2020). We computed two types of similarity: First, we computed *Intra-Model Similarity*, i.e., pairwise cosine similarity among all responses from the same source (i.e., model or humans) for a given question

[3]Note that sources of each model can be found in Table 3 in the Appendix.

[4]Note that we left out all responses containing the default tag `<EDIT-NO-RESPONSE>`

| Model | Parameter Size | Architecture | Domain |
|---|---|---|---|
| flan-t5-base (standard) | 250 M | Encoder-Decoder | general |
| flan-t5-base (medical) | 250 M | Encoder-Decoder | biomedical |
| biogpt | 347 M | Decoder-Only | biomedical |
| BioGPT-MedText | 347 M | Decoder-Only | biomedical |
| Llama-3.2-1B-Instruct | 1 B | Decoder-Only | general |
| Bio-Medical-Llama-3-2-1B-CoT-012025 | 1 B | Decoder-Only | biomedical |
| Llama-3.2-3B-Instruct | 3 B | Decoder-Only | general |
| Phi-4-mini-instruct | 3.8 B | Decoder-Only | general |
| gemma-3-4b-it | 4 B | Decoder-Only | general |
| bloom-6b4-clp-german | 6 B | Decoder-Only | general |
| Qwen2.5-7B-Instruct | 7 B | Decoder-Only | general |
| Qwen-UMLS-7B-Instruct | 7 B | Decoder-Only | biomedical |
| Mistral-7B-Instruct-v0.1 | 7 B | Decoder-Only | general |
| BioMistral-7B | 7 B | Decoder-Only | biomedical |
| Ministral-8B-Instruct-2410 | 8 B | Decoder-Only | general |

Table 2: Overview of models used for synthetic data generation.

to quantify internal variability. Second, we calculated *Inter-Model Similarity*, where we used cosine similarity between response centroids, i.e., the *average response*, to compare models with each other. In addition, we computed how distant individual model responses were to their own model centroids and to the respective human centroids.

#### 4.2.2 Results

The average intra-model cosine similarity aggregated over all questions is 0.43±0.06 for humans, 0.44±0.13 for the biomedical LLMs, and 0.52±0.18 for the general LLMs.

A two-way ANOVA (Girden, 1992) was conducted to examine the effects of *domain* and *question type* on *intra-model diversity*, operationalized as the pairwise cosine similarity between single responses. The analysis revealed significant main effects of *domain* ($F(2, 13215) = 673.69$, $p < .001$) and *question type* ($F(2, 13215) = 12.17$, $p < .001$), as well as a significant interaction effect between the two ($F(4, 13215) = 40.43$, $p < .001$)).

A post-hoc Turkey HSD test (Tukey, 1949) showed that for CQ, general-domain LLMs were significantly less diverse than biomedical models (mean difference (md) = 0.084, 95 % CI = [0.076, 0.092], $p < .001$) and humans (md = 0.128, $p < .001$). Biomedical models also produced more similar responses than humans, though with a smaller effect (md = 0.044, p = .029). For PQ, general LLMs again exhibited lower diversity than biomedical models (md = 0.091, $p < .001$) and humans (md = 0.088, $p < .001$). In contrast, no significant difference was found between humans and biomedical models (p = .77). For WhQ, general LLMs remained less variable than both biomedical models (md = 0.057, $p < .001$) and humans (md = 0.032, $p < .001$). Interestingly, humans produced less diverse responses than biomedical models in this condition (mean difference = 0.024, p = .0017).

The analysis of the inter-model similarity shows that the average distance, i.e., the inversed cosine similarity of the synthetic responses by the biomedical LLMs to the centroid human responses is slightly higher (avg = 0.64±0.15) than the generations from the general models (avg = 0.60±0.15). A two-way ANOVA showed significant effects of the *model's domain* ($F(1,8694) = 134.63$, $p < .001$), the *question type* ($F(2, 8694) = 6.50$, $p < 0.01$), and the interaction between both variables ($F(2, 8694) = 3.63$, p = .027) when predicting the *distance to the human centroid*. Post-hoc pairwise Turkey-adjusted comparisons revealed that for all question types, the biomedical centroid responses are significantly less distant from the human centroid, while for CQ (md = 0.055, $p < .001$) and PQ (md = 0.037, $p < .001$), the effect is large. WhQ reveal a similar trend but with a lower effect (md = 0.029, p = .004).

To account for similarity relations between unique models, we selected for each LLM the top ten percent of most similar counterparts based on the centroid response similarity. The responses of `Phi-4-mini-instruct` are most similar to the centroid human responses (avg = 0.64, 95 % CI [0.61, 0.67]). The highest similarity between any centroid responses can be found between `Ministral-8B-Instruct-2410` and `Qwen2.5-7B-Instruct` (avg = 0.84, 95 % CI [0.82, 0.86]), while the lowest top-similarity between two models is found with `BioGPT-MedText` and `Bio-Medical-Llama-3-2-1B-CoT-012025` (avg = 0.54, 95 % CI [0.51, 0.57]).

Figure 1 illustrates a similarity graph where each node represents the centroid responses of

a model. An edge is added between two nodes if one of them represents the most similar centroid response of the other node among the top ten percent. The models `Phi-4-mini-instruct` and `Qwen2.5-7B-Instruct` produced four times each the most similar centroid responses to other models, indicating a large overlap in the generated responses. The figure shows two similarity-islands: Both `flanT5` models seemed to produce very similar responses, regardless of the domain, and, the small-sized LLMs `BioGPT-MedText` and `Llama-3.2-1B-Instruct` produced a large overlap, too.

#### 4.2.3 Interim Discussion

The analysis of the responses showed that intra-model similarity is lowest for humans and highest for general LLMs, suggesting more diversity in the human responses. Further dividing the scores based on the question types shows that WhQ evoked the least diverse responses from humans. This is a counterintuitive finding if we ground our knowledge on the semantic structure of questions, where WhQ are known to denote a much larger set of possible responses than PQ. Crucially, this trend does not apply to the synthetic responses. Nevertheless, the intra-model similarity of the synthetic responses largely corresponds to the diversity of human responses, which could be understood as a hint that models actually meet the requirements to generate diverse enough answers to standardized medical questions.

The assessment of the inter-model similarity revealed clusters among the most similar responses of certain models. The encoder-decoder models, as well as the smaller models `Llama-3.2-1B-Instruct` and `BioGPT-MedText`, were found to be *similarity islands*, which are detached from the other models. The islands might be due to high redundancy in the generated responses. The most similar model to the human responses was `Phi-4-mini-instruct`, which also appeared to be most similar to the largest number of other models. Thus, `Phi-4-mini-instruct` might be seen as a very representative synthetic response generator.

### 4.3 LLM-as-a-Judge

The previous two evaluations focused on the comparison of the structural and semantic properties of the human-produced and synthetic responses to the anamnesis questions. In order to assess whether the responses are actually meaningful or appropriate, we conducted an LLM-as-a-judge experiment (cf. Zheng et al., 2023).

#### 4.3.1 Methods

We evaluated all responses with five pretrained LLMs from the same pool of models as the ones that generated the responses. In particular, we selected `Qwen2.5-7B-Instruct`, `Qwen-UMLS-7B-Instruct`, `Bio-Medical-Llama-3-2-1B-CoT-012025`, `Llama-3.2-3B-Instruct`, and `gemma-3-4b-it` as judges. Models were loaded via the vLLM Python API (Kwon et al., 2023). For each model, we retrieved its default SamplingParams and constrained the response length to 70 tokens max.

To elicit appropriateness ratings, we designed a system–user prompt template in English. The system message instructed the model to rate each response on a Likert scale from 1 (not appropriate) to 5 (very appropriate), emphasizing naturalness, coherence, and contextual fit, and to reply with a single digit only. A comparison between that template and an identical German prompt, a prompt that asks the model to justify its response, and one that required the model to produce three ratings for each of the aforementioned rating criteria showed no relevant differences in the judgements.

The prompts were dynamically generated for each question-response pair. Non-numeric responses, or those that exceeded the borders of the Likert scale, received the default tag `<EDIT-NO-JUDGEMENT>`.

An evaluation of the inter-rater-agreement between our LLM-judges showed a very low Fleiss' $\kappa$ of 0.08. Therefore, we excluded all neutral scores from the dataset and removed the ratings from the two least agreeing LLM-judges, i.e., both Llama models. Furthermore, we dichotomized the scores into inappropriate (1 & 2) and appropriate responses (4 & 5). This trimmed, binary analysis yielded a substantially higher Fleiss' $\kappa$ of 0.72, indicating substantial agreement on definitive adequacy judgments.

#### 4.3.2 Results

The aggregated binary scores of the three LLM judges showcase that human responses were consistently rated higher, thus more appropriate, than synthetic responses throughout all question types. Figure 2 illustrates the proportion of high and low ratings for human and synthetic data points grouped

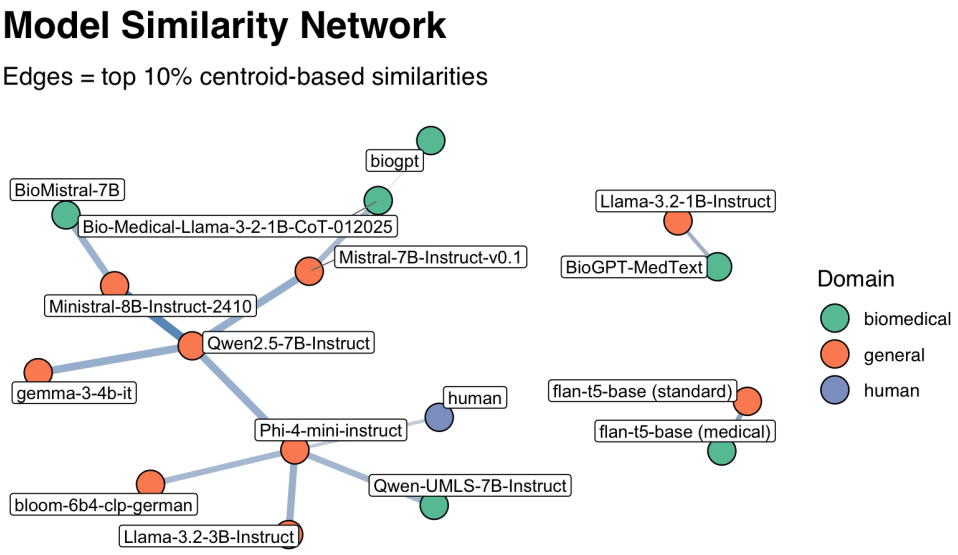


Figure 1: Network graph connecting the most similar centroid responses.

by the judgement models.

We fitted a logistic regression model to investigate how the binary rating outcome is influenced by the fact that a model or a human produced the responses (*human-ness*), *question type*, and the *LLM judge*. Interaction effects between the *human-ness* and *question type* were included to assess whether the effect of human versus non-human responses varies by question type. The model showed excellent overall predictive discrimination (AUC = 0.98). All predictors were statistically significant, reflecting the large sample size (~1.4M observations). The primary factor distinguishing ratings was whether the response was generated by a human or a machine (Odds Ratio (OR) for synthetic responses = $4.5 \cdot^{-5}$, 95 % CI: $[4.0 \cdot 10^{-5}, 5.0 \cdot 10^{-5}]$, $p < .001$). Non-human responses thus had substantially lower odds of being rated positively compared to human responses. Significant interaction effects indicated that the magnitude of the human advantage varied by question type. For instance, compared to the reference category (CQ), PQ showed a moderated reduction of human advantage (interaction OR = 0.57, 95 % CI: [0.49, 0.66], $p < .001$), whereas WhQ increased the human advantage (interaction OR = 32.3, 95 % CI: [27.0, 38.6], $p < .001$). Additionally, judgment significantly influenced the rating outcomes, although with smaller effect sizes. For instance, responses judged by `Qwen2.5-7B-Instruct` had considerably lower odds of receiving positive ratings compared to the baseline (OR = 0.13, 95 % CI: [0.13, 0.14], $p < .001$), whereas `Qwen-UMLS-7B-Instruct` had modestly lower odds (OR = 0.86, 95 % CI: [0.84, 0.88], $p < .001$).

#### 4.3.3 Interim Discussion

The rating experiment revealed that the LLM judges scored human-produced responses significantly higher than synthetic ones. WhQ even increase the human advantage, possibly due to the semantic nature of this question type, i.e., usually they require knowledge about the space of possible responses. This trend is not influenced by more critical LLM judges.

## 5 General Discussion

The driving question behind the linguistic-structural, semantic, and adequacy evaluation of the GerMedIQ corpus and its augmented counterparts was to identify whether small to medium-sized, open-weight LLMs serve as reliable synthetic data generators. Structurally, the chosen LLMs produced much longer responses than humans, and general LLMs produced the longest responses. It is difficult to account for whether longer or shorter responses are better in this case; however, a reason for the LLMs to produce longer responses may be the general training strategy of those models: LLMs are usually not designed to produce short

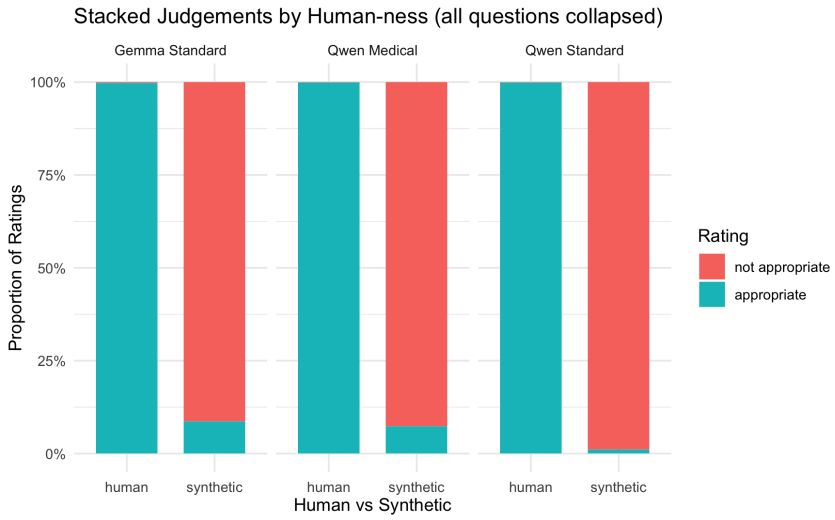


Figure 2: Proportion of responses judged as either appropriate or not appropriate by the three LLM judges, grouped by

and concise responses, but rather detailed and often also step-by-step justifications (cf. Zhang et al., 2024). Albeit the response length, the models managed to produce almost identical n-grams as the humans, indicating the ability to meet the major structure of the task: Formulating a grammatically well-formed response. This minimum requirement seemed to have worked out quite well.

From a distributional semantics point of view, the LLMs produced responses with a slightly lower diversity than human responses, although the general tendency is comparable to human performance. Looking at the top inter-model similarities revealed that especially smaller models produced highly similar responses, suggesting redundancies. On the other hand, `Phi-4-mini-instruct` could be identified as a representative response generator: Its responses were the most similar to three other models and the human responses. Overall, the semantic analysis showed that most LLMs generally managed to produce responses that are comparably diverse and semantically related to human responses.

The LLM-as-a-judge study clearly draws a different picture. All judges agreed throughout the question types and model domains that humans always gave more adequate responses. These ratings, although not grounded by human judgements, can be seen as a strong hint towards a quality gap. WhQ even increased that gap, indicating that LLMs struggle with the semantic requirements of that question type.

Altogether, the experiments have shown that the usage of LLMs for data augmentation in the context of German clinical language has to be done with care. A life-cycle for synthetic textual data or a human-in-the-loop approach might be important to consider before further processing the data (cf. Long et al., 2024; Liu et al., 2024).

## 6 Conclusion

We release a novel simulated medical anamnesis interview question dataset, 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 humans.

Moreover, we could show that small to medium-sized LLMs should only be leveraged carefully as synthetic data generators in narrow domains and non-English contexts. While most LLMs managed to meet the structure of human responses, some models showed weak semantic similarity values compared to the human data. The rating study clearly rejected LLMs as adequate response augmentation machines.

Future research should investigate further whether LLMs behave similarly in other non-English contexts, perhaps allowing a more diverse set of LLMs, including larger models. In order to ground the rating experiment, a comparable human rating study appears to be a fruitful extension of our research.

## 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 made 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 used small to medium-sized LLMs with a similar prompt for all models. Leveraging larger LLMs and finding optimal prompts for individual models might lead to different 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, and participants were informed on how the data would be used. All our experiments were conducted with open-source libraries, which received due citations. The experiment is in line with the ethical regulations of the REDACTED.

## 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.

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*.

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.

Ellen R. Girden. 1992. *ANOVA: Repeated measures*. 84. Sage.

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. 2024. 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. *arXiv preprint*. ArXiv:2412.00230 [cs].

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).

Matthew Honnibal, Ines Montani, Sophie van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-Strength Natural Language Processing in Python.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei 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.

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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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.

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, 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.

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.

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.

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.

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.

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

John W. Tukey. 1949. Comparing Individual Means in the Analysis of Variance. *Biometrics*, 5(2):99.

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.

Yusen Zhang, Sarkar Snigdha Sarathi Das, and Rui Zhang. 2024. Verbosity $\neq$ Veracity: Demystify Verbosity Compensation Behavior of Large Language Models. *Preprint*, arXiv:2411.07858.

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

Table 3 lists all used LLMs with their corresponding research papers, if applicable. If no research paper has been released, the Huggingface URLs are given instead.

| Model | Source |
|---|---|
| flan-t5-base (standard) | Chung et al. (2022) |
| flan-t5-base (medical) | https://huggingface.co/CLARA-MeD/flan-t5-base |
| biogpt | Luo et al. (2022) |
| BioGPT-MedText | https://huggingface.co/AventIQ-AI/BioGPT-MedText |
| Llama-3.2-1B-Instruct | https://huggingface.co/meta-llama/Llama-3.2-1B |
| Bio-Medical-Llama-3-2-1B-CoT | https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-2-1B-CoT-012025 |
| Llama-3.2-3B-Instruct | https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct |
| Phi-4-mini-instruct | https://huggingface.co/microsoft/Phi-4-mini-instruct |
| gemma-3-4b-it | https://huggingface.co/google/gemma-3-4b-it |
| bloom-6b4-clp-german | Ostendorff and Rehm (2023) |
| Qwen2.5-7B-Instruct | Yang et al. (2024); Team (2024) |
| Qwen-UMLS-7B-Instruct | https://huggingface.co/prithivMLmods/Qwen-UMLS-7B-Instruct |
| Mistral-7B-Instruct-v0.1 | https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1 |
| BioMistral-7B | Labrak et al. (2024) |
| Ministral-8B-Instruct-2410 | https://huggingface.co/mistralai/Ministral-8B-Instruct-2410 |

Table 3: LLMs and their corresponding sources.