Collaboration across Disciplines for the Health of People, Animals and Ecosystems
L. Stoicu-Tivadar et al. (Eds.)

doi:10.3233/SHTI241099

Generating Synthetic Healthcare Dialogues in Emergency Medicine Using Large Language Models

Denis MOSER[a], Matthias BENDER[a] and Murat SARIYAR[a,1]
[a]*Bern University of Applied Sciences, Switzerland*
ORCiD ID: Murat Sariyar https://orcid.org/0000-0003-3432-2860

Abstract. Natural Language Processing (NLP) has shown promise in fields like radiology for converting unstructured into structured data, but acquiring suitable datasets poses several challenges, including privacy concerns. Specifically, we aim to utilize Large Language Models (LLMs) to extract medical information from dialogues between ambulance staff and patients to populate emergency protocol forms. However, we currently lack dialogues with known content that can serve as a gold standard for an evaluation. We designed a pipeline using the quantized LLM "Zephyr-7b-beta" for initial dialogue generation, followed by refinement and translation using OpenAI's GPT-4 Turbo. The MIMIC-IV database provided relevant medical data. The evaluation involved accuracy assessment via Retrieval-Augmented Generation (RAG) and sentiment analysis using multilingual models. Initial results showed a high accuracy of 94% with "Zephyr-7b-beta," slightly decreasing to 87% after refinement with GPT-4 Turbo. Sentiment analysis indicated a qualitative shift towards more positive sentiment post-refinement. These findings highlight the potential and challenges of using LLMs for generating synthetic medical dialogues, informing future NLP system development in healthcare.

Keywords. Natural Language Processing, Large Language Model (LLM), Synthetic data generation, Emergency medical services, Retrieval-Augmented Generation (RAG), Sentiment analysis.

1. Introduction

Data is fundamental to the healthcare system, both in its structured and unstructured forms. Healthcare Information Systems (HISs), such as Electronic Health Records (EHRs), rely on structured data to support healthcare professionals (HCPs) in delivering effective services. Structured data also enables quantitative analyses via statistical and machine learning methods [1]. Creating structured health data often requires significant manual effort, making the process time-consuming and labor-intensive. Simplifying the conversion of unstructured data into structured formats is essential in the extraction process [2,3]. For instance, the value of Natural Language Processing (NLP) in radiology has been recognized, particularly in converting unstructured data from radiology reports into structured formats [3]. The development of such NLP systems relies on suitable

[1] Corresponding Author: Murat Sariyar, Bern University of Applied Sciences, Quellgasse 21, CH2502 Biel/Bienne, Switzerland; E-mail: murat.sariyar@bfh.ch.

training data, including medically relevant information and annotations. Due to privacy and security concerns, it is often challenging to acquire appropriate datasets.

Synthetic data generation offers an alternative to the acquisition of real-world data. Our specific use case involves using Large Language Models (LLMs) to extract medical information from dialogues between ambulance staff and patients to populate emergency service protocol forms. However, we lack sufficient real-world dialogues to develop and evaluate the NLP pipeline. Additionally, existing datasets and approaches are not well-aligned with our context. For instance, the NoteChat cooperative multi-agent framework, which uses LLMs to generate patient-physician dialogues, requires additional training to generate dialogues involving other healthcare scenarios and personnel [4]. This is why we propose a novel NLP pipeline designed to generate medical dialogues across various contexts, including emergency medicine within a German context.

The primary research question guiding this work is: How can we generate near-realistic medical dialogues that effectively simulate interactions between ambulance staff and patients, while ensuring the inclusion of relevant and known medical content? Accurate and realistic dialogue generation is crucial for developing and testing advanced text extraction algorithms used in emergency medical systems. In this context, a "dialogue" refers to a simulated conversation between ambulance staff and patients or other medical personnel. For instance, a dialogue might involve a patient reporting chest pain and the ambulance crew asking detailed questions about the onset, location, and nature of the pain, followed by appropriate medical interventions. Such dialogues provide a controlled environment for assessing how well these systems can interpret and respond to real-world scenarios. In the following, we show how to design and evaluate a pipeline based on locally employed LLMs for generating synthetic German dialogues in an emergency dispatch scenario.

2. Methods

This section describes the pipeline developed for generating and expanding dialogues between ambulance staff and patients using a combination of LLMs, prompt engineering, and database interactions, as well as the subsequent evaluation of the generated dialogues. Our pipeline consists of three stages: initial dialogue generation, refinement and translation into German, and finally evaluation. As data source, the ED module of the MIMIC-IV database, a comprehensive and publicly accessible repository containing anonymized patient records, was used. We stored the data in a MongoDB database. The aim was to leverage real-world medical information and to minimize constellations where LLMs tend to produce hallucinations or inaccuracies [5]. Our goal is to create near-realistic dialogues that serve as a gold standard for evaluation. These dialogues simulate various medical scenarios pertinent to emergency protocols, reflecting authentic interactions between ambulance staff and patients. By issuing precise instructions to the language model, we ensure the dialogues encompass reciprocal exchanges and cover all relevant medical considerations. The system is designed to generate dialogues that are not only contextually accurate but also diverse in content to cover a broad range of medical scenarios. The MIMIC dataset supplies the necessary medical information. Although the dialogues may seem somewhat formal, they are designed to optimize the assessment of text extraction algorithms in future research. Examples of these dialogues will be provided in a subsequent section due to space constraints.

The initial dialogue generation is performed using the quantized LLM "Zephyr-7b-beta" model from the Hugging Face transformers library. Dialogues are generated considering four phases: initial ambulance interaction, triage and first measurements of vital signs, medication and second measurements of vital signs, and hospital arrival, forming a coherent story arc. For the different phases, prompt templates are developed, using real-world domain knowledge as input for general instructions (e.g., medical procedures to be undertaken) and placeholders for corresponding MIMIC-IV data items as specifics. After each prompt, the responses from the LLM are appended to form a coherent dialogue text. The process is iterated 100 times through the MongoDB collection, storing the resulting dialogues for further processing.

In the refine stage, the dialogues are expanded and translated using OpenAI's GPT-4 Turbo, as the local LLMs struggle with the naturalness of speech and translation accuracy. The process begins by retrieving initial dialogues from MongoDB and expanding them using a hierarchical technique from OpenCredo [6]. This method involves breaking down dialogues into thematic sections and creating detailed outlines with summaries to guide realistic interactions. Prompt templates, including placeholders for the outlines and summaries, are created to ensure comprehensive coverage of patient and HCP communication and the medical information (e.g., diagnosis and vital signs). The expansion is performed iteratively. For each initial dialogue, outlines are generated, summarizing important information. The LLM is then prompted with a template instructing it to generate dialogue text based on the context, including the outline summary and, if available, the summary of the previous expanded dialogue to maintain continuity, and when appropriate, adding relevant medical treatment procedures. The model then returned the newly expanded dialogue in German. This process is repeated for the 100 initial dialogues generated by the local LLM.

In the evaluation stage, we measure accuracy, defined as the presence of relevant medical information incorporated at the beginning of the dialogue generation pipeline. For each set of generated dialogues, we use a Retrieval-Augmented Generation (RAG) system to assess this accuracy. The RAG pipeline operates as follows: (i) it retrieves dialogues from the MongoDB database, preprocesses them into chunks, and generates embeddings using the "intfloat/multilingual-e5-large-instruct" model and the FAISS vector store. Prompt templates are designed to query specific MIMIC-IV data items, such as "Are there heart rates mentioned with values like ‘heart rate is 80’?" (ii) After executing the prompts, the RAG system, utilizing the multilingual "TwT-6/cr-model-v1" model, retrieves relevant information or responds with "No information available in the context". The results are compiled into a structured JSON format for further evaluation. (iii) For final scoring, we use the sentence transformer model 'BAAI/bge-m3' to generate embeddings for the questions and the RAG-generated responses. We then calculate cosine similarity to measure the similarity between the provided information and the expected medical data. The similarity threshold for determining a match was established based on the highest similarity score observed among the "No information" responses. Accuracy is measured by calculating the proportion of true matches in the dialogues.

In addition to the accuracy computation, we perform a sentiment analysis on the generated dialogues to evaluate the emotional tone in the initial and refined dialogues, aiming to obtain a qualitative understanding and compare differences between the local LLM and the OpenAI model. The dialogues are fetched from the MongoDB database, cleaned, and split into sentences. Using the “distilbert-base-multilingual-cased-sentiments-student” model, sentiment scores for negative, neutral, and positive sentiments are generated for each sentence. The highest sentiment score for each sentence is identified to determine the majority sentiment (negative, neutral, or positive). We then calculate the overall percentage distribution of these majority sentiments to understand the prevalent emotional tone in the dialogues.

3. Results

A total of 200 dialogues were generated, 100 initially with the local model, and then refined using OpenAI. First, the accuracy of the dialogues was determined, for the initial dialogues generated using "Zephyr-7b-beta", the similarity cutoff value was set to 0.43, determined by the maximum similarity score of negative responses. Out of 1288 instances (RAG retrievals, one for each injected MIMIC-IV data item), 1206 had similarity scores greater than the cutoff, resulting in an accuracy of 94%. For the refined dialogues using GPT-4 Turbo, the cutoff value was set to 0.44. Out of 1288 instances, 1116 had similarity scores greater than the cutoff, resulting in an accuracy of 87%, leading to the conclusion, as shown in Table 1, that further text refinement had led to information loss. We manually reviewed the results and did not find any false positives, so their occurrence is likely to be very low.

Table 1. Summary of the accuracies of the generated dialogues, showing that the further refinement with GPT-4 Turbo and hierarchical expansion resulted in a slight loss of accuracy compared to the initial dialogues generated using Zephyr-7b-beta.

Dialogue Set	Model	Cutoff	Total Instances	Instances > Cutoff	Accuracy (%)
Initial	Zephyr-7b-beta	0.43	1288	1206	94
Refined	GPT-4 Turbo	0.44	1288	1116	87

Table 2 presents the sentiment analysis of the dialogues. Initially generated dialogues exhibited a predominantly negative sentiment, with 68% of the dialogues having the highest sentiment score categorized as negative. In contrast, the refined dialogues generated using GPT-4 Turbo demonstrated a reduced negative sentiment distribution, with only 62% classified as negative. This shift indicates that the refinement introduced a qualitative shift towards more positive sentiment.

Table 2. Sentiment distribution in generated dialogues, indicating a shift toward more positive sentiment in each sentence after using GPT-4 Turbo for refining the initial dialogue text.

Overall Sentiment	Initial Dialogues (Zephyr-7b-beta) (%)	Refined Dialogues (GPT-4 Turbo) (%)
Negative	68	62
Positive	26	34
Neutral	6	4

4. Discussion and Conclusions

The results of our study highlight the potential and challenges of using LLMs for generating synthetic medical dialogues, specifically in the context of emergency medicine. Our initial experiments with the "Zephyr-7b-beta" model demonstrated a high accuracy in the inclusion of relevant medical information from the MIMIC-IV database, with a 94% accuracy rate. This underscores the capability of locally pre-trained LLMs to generate detailed and contextually appropriate medical dialogues. However, the refinement process using GPT-4 Turbo, while improving the naturalness and fluency of the dialogues, led to a slight decrease in the accuracy of the medical information retained, dropping to 87%. This trade-off between linguistic quality and informational accuracy is a critical consideration in the development of NLP systems for healthcare applications. The loss of accuracy suggests that GPT-4 Turbo may introduce subtle inaccuracies or omissions in the medical content, which could impact the reliability of the generated data for clinical use. The sentiment analysis further confirmed that by revealing a qualitative shift in the emotional tone of the GPT-4 generated dialogues. The initial dialogues, predominantly negative in sentiment, were transformed to have a more balanced distribution of sentiment, with an increase in positive sentiments.

Despite the promising results, there are limitations to our approach that warrant further investigation. The reliance on synthetic data for training and evaluation poses inherent challenges, including the risk of overfitting to the synthetic examples and potential biases introduced during the data generation process. Moreover, while our use case focused on emergency medicine, the applicability of the developed pipeline to other medical contexts needs to be explored to ensure its generalizability and robustness.

In conclusion, our study demonstrates the feasibility of utilizing LLMs to generate synthetic medical dialogues through an advanced RAG pipeline. These dialogues are valuable for encapsulating relevant medical information, thereby serving as a benchmark by establishing a gold standard. The primary goal moving forward is to produce dialogues at scale to support the evaluation of text extraction algorithms and to fine-tune LLMs, particularly with German texts. Furthermore, we will conduct a comparative analysis between synthetic and real dialogues to evaluate potential biases.

Acknowledgments: This study was funded by the Swiss Inventus foundation (grant number 058/2024).

References

[1] Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, et al. Challenges and opportunities beyond structured data in analysis of electronic health records. WIREs Computational Statistics 2021;13:e1549. https://doi.org/10.1002/wics.1549.

[2] Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. J Am Med Inform Assoc 2011;18:181–6. https://doi.org/10.1136/jamia.2010.007237.

[3] Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, et al. A systematic review of natural language processing applied to radiology reports. BMC Medical Informatics and Decision Making 2021;21:179. https://doi.org/10.1186/s12911-021-01533-7.

[4] Wang J, Yao Z, Yang Z, Zhou H, Li R, Wang X, et al. NoteChat: A Dataset of Synthetic Doctor-Patient Conversations Conditioned on Clinical Notes 2023. https://doi.org/10.48550/arXiv.2310.15959.

[5] Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models 2023. https://doi.org/10.48550/arXiv.2309.01219

[6] Nutall G. How to use LLMs to Generate Coherent Long-Form Content using Hierarchical Expansion. https://opencredo.com/blogs/how-to-use-llms-to-generate-coherent-long-form-content-using-hierarchical-expansion/ (accessed June 25, 2024).