Optimizing data extraction: harnessing RAG and LLMs for German medical documents

Yingding Wang, Simon Leutner, Michael Ingrisch, Christoph Klein, Ludwig Christian Hinske, Katharina Danhauser

Angaben zur Veröffentlichung / Publication details:

Wang, Yingding, Simon Leutner, Michael Ingrisch, Christoph Klein, Ludwig Christian Hinske, and Katharina Danhauser. 2024. "Optimizing data extraction: harnessing RAG and LLMs for German medical documents." In Digital health and informatics innovations for sustainable health care systems: proceedings of MIE 2024, edited by John Mantas, Arie Hasman, George Demiris, Kaija Saranto, Michael Marschollek, Theodoros N. Arvanitis, Ivana Ognjanović, et al., 949–50. Amsterdam: IOS Press. https://doi.org/10.3233/shti240567.

Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240567

Optimizing Data Extraction: Harnessing RAG and LLMs for German Medical Documents

Yingding WANG[a], Simon LEUTNER[b], Michael INGRISCH[c], Christoph KLEIN[a], Ludwig Christian HINSKE[d] and Katharina DANHAUSER[a,1]
[a] *Department of Pediatrics, Dr. von Hauner Children's Hospital, University Hospital, LMU Munich, Munich, Germany*
[b] *Medical Technology and IT (MIT), University Hospital, LMU Munich, Munich, Germany*
[c] *Department of Radiology, University Hospital, LMU Munich, Munich, Germany*
[d] *Institute for Digital Medicine, University Hospital Augsburg, Augsburg, Germany*
ORCiD ID: Michael Ingrisch https://orcid.org/0000-0003-0268-9078
Christoph Klein https://orcid.org/0000-0003-0956-0445
Ludwig Christian Hinske https://orcid.org/0000-0001-7273-5899
Katharina Danhauser https://orcid.org/0009-0000-7734-2293

Abstract. In the field of medical data analysis, converting unstructured text documents into a structured format suitable for further use is a significant challenge. This study introduces an automated local deployed data privacy secure pipeline that uses open-source Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) architecture to convert medical German language documents with sensitive health-related information into a structured format. Testing on a proprietary dataset of 800 unstructured original medical reports demonstrated an accuracy of up to 90% in data extraction of the pipeline compared to data extracted manually by physicians and medical students. This highlights the pipeline's potential as a valuable tool for efficiently extracting relevant data from unstructured sources.

Keywords. OSS-LLM, Data extraction, RAG, German, Real-life medical reports

1. Introduction

In medical informatics, processing unstructured text data while ensuring data protection and confidentiality is a major challenge [1]. This study introduces a solution using open-source Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) architecture [2] to efficiently transform this data into a structured format. By evaluating the automated pipeline on numerous medical documents, we aim to showcase its effectiveness in extracting and structuring data, as a step forward in health informatics.

[1] Corresponding Author: Katharina Danhauser; E-mail: kdanhaus@med.lmu.de.

2. Methods

Our method centers on a local deployed automated pipeline using different OSS-LLMs with RAG architecture to process German medical documents, extract and structure important data, and avoid external data transfer to ensure privacy. The pipeline works in stages: 1) it inputs unstructured reports, 2) translates them from German to English using an OSS translation model, 3) uses RAG to identify and retrieve information, and 4) then converts these snippets into structured data for downstream use. To assess the accuracy of the data extraction process we compared the extracted data with manually extracted information by medical experts, which served as the ground truth.

3. Results

We evaluated real life private datasets of 800 unstructured medical reports to assess effectiveness of the pipeline. Figure 1 displays initial results for weight and age.

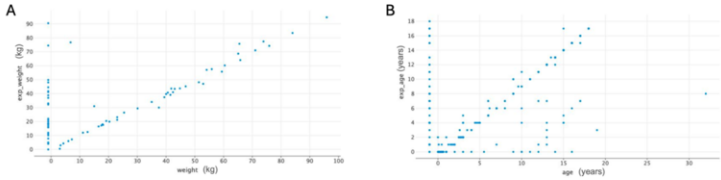

Figure 1. Residual analyses for weight (A) and age (B) extraction from German medical reports. Y-axis displays the expected values, and X-axis shows the automatically extracted information from the texts.

4. Discussion

First results show that an automated pipeline can efficiently convert unstructured medical documents into structured data with up to 90% accuracy. This highlights the effectiveness of using OSS-LLMs with RAG for high accuracy data extraction while emphasizing strong data protection. Challenges such as improving the model's handling of various document formats and medical terms are part of future research.

5. Conclusions

The initial result indicates that the automated pipeline is efficient for converting unstructured medical data. By initially translating the text, the approach can be tailored to accommodate various languages, potentially broadening its applicability globally.

References

[1] Mohamed Yassine Landolsi, Lobna Hlaoua, and Lotfi Ben Romdhane. Information extraction from electronic medical documents: state of the art and future research directions. Knowl Inf Syst. 2023;65(2):463-516. doi: 10.1007/s10115-022-01779-1.

[2] Wang C., Ong J., Wang C., Ong H., Cheng R., Ong D. Potential for GPT Technology to Optimize Future Clinical Decision-Making Using Retrieval-Augmented Generation. Ann Biomed Eng. 2023 Aug 2. doi: 10.1007/s10439-023-03327-6.