SYNTHETIC-BASED RETRIEVAL OF PATIENT MEDICAL DATA

Rinat Mullahmetov
Department of Personalization
Zvuk
Innopolis, Russia
r.mullahmetov@innopolis.ru

Ilya Pershin
Research Center of the Artificial Intelligence Institute
Innopolis University
Innopolis, Russia
i.pershin@innopolis.ru

ABSTRACT

Medical retrieval systems play a crucial role in facilitating an accurate and efficient diagnosis by allowing physicians to access relevant radiological reports and patient descriptions. However, the development of such systems is often hindered by the limited availability of high-quality labeled data due to privacy concerns and data scarcity. In this work, we propose an approach to address this challenge by using synthetic data generation using Large Language Models (LLMs). Our experiments show that synthetic data is useful for improving retrieval performance in various tasks, both in training modes entirely on synthetic data and in a mixed-with-real-data mode.

1 INTRODUCTION

Retrieval-based Clinical Decision Support (ReCDS) systems play a crucial role in modern healthcare by providing clinicians with similar patient cases, thereby facilitating evidence-based decision making Zhao et al. (2023). However, the development of effective ReCDS systems has been hindered by the scarcity of diverse, large-scale patient datasets and the challenges associated with manual data annotation Ahmed et al. (2023).

Recent advances in synthetic data generation present a promising solution to these limitations. Using large-language models (LLMs), researchers can generate high-quality synthetic data that captures the complexity and diversity of real-world medical cases Wang et al. (2024). This approach not only addresses privacy concerns associated with real patient data but also enables the creation of more comprehensive training datasets.

In this study, we address two key retrieval tasks in the medical domain: query-to-patient data retrieval and patient-to-patient data retrieval. To achieve this, we use LLM to generate synthetic data, which is then used to fine-tune text embedding model. We involve radiologists to generate the prompt and validate the generation results. Our experiments demonstrate that synthetic data significantly improved performance in medical retrieval tasks.

2 METHODOLOGY

Our methodology is based on the approach presented by Wang et al. (2024), adapted for medical data. We work with patient information containing diagnostic lists and radiological text descriptions. To generate synthetic data, we employ a large language model (LLM) that creates positive and negative queries based on this patient information. The LLM takes the patient's diagnostic details and radiological description as input to produce relevant positive queries and intentionally irrelevant negative queries for each patient record. We engaged 3 radiologists to collaborate on creating realistic prompts that were similar to the needs of clinical practice.

We then use these generated query pairs to train an embedding model through contrastive learning in a report-to-query setting. For evaluation, we construct a test dataset by selecting patients with single-pathology cases. Positive samples consist of patients with the same specific pathology, while negative samples include patients with any other pathologies except the selected one. We evaluate

Table 1: Retrieval R-Precision metric on two test datasets (Qwery-to-report and Report-to-report) for different approaches to using data for fine-tuning.

Data for fine-tune	R-Prec ↑, Qwery-to-report	R-Prec ↑, Report-to-report
Stella pretrain	0.653 ± 0.02	0.608 ± 0.006
Synth. queries	0.722 ± 0.015	0.696 ± 0.005
Original reports	0.78 ± 0.014	0.86 ± 0.004
Synth. queries and original reports	$\mathbf{0.798 \pm 0.015}$	$\mathbf{0.891 \pm 0.004}$

our model in two distinct modes: query-to-report and report-to-report. In the query-to-report mode, we generate synthetic queries using LLM based on patient descriptions for selected patients, while in the report-to-report mode, we use original radiological reports without modification. Radiologists validated all generated queries.

Furthermore, we compare three training approaches: query-to-report using only synthetic queries and original reports, report-to-report using only original radiological reports, and a mixed mode that combines both synthetic queries and genuine radiological reports during training.

3 Experiments and results

Our experiments were conducted using the publicly available MIMIC-CXR dataset Johnson et al. (2024), which is a comprehensive database of chest X-ray images. The dataset includes radiological descriptions, metadata about patients (such as the presence of pathologies, orientation of the X-ray, and patient positioning), and corresponding chest X-ray images. For our experiments, we selected 5000 samples for training and 10000 samples for testing. To generate synthetic queries, we utilized the Qwen2.5-72B-Instruct-GPTQ-Int4-instruct model Yang et al. (2024). We used the following system prompt to generate positive and negative samples:

> You are a language model trained to assist in generating data for contrastive learning of medical text embeddings. Your task is to create positive and negative queries for a given medical report.
>
> A positive query should closely align with the medical report. It can focus on either the complete report or specific key findings. Vary the length, level of detail, and phrasing to ensure diversity. A negative query should introduce clear contradictions, irrelevant findings, or completely unrelated contexts. Ensure that negative queries differ significantly in focus, while remaining plausible as medical queries. Rules for Variability:
>
> Vary the length: Create some queries that are short summaries and others that are detailed descriptions. Vary the specificity: Some queries should focus on a subset of the report (e.g., ribs or lungs), while others cover the entire report. Vary the style: Use questions, statements, or even keyword-based queries to create variation.
>
> * Use english
> * Use different sized formulations
> * Output format: {'positive':", 'negative': "}

We used the following system prompt to generate the query for the report:

> Task:
>
> Based on the patient's medical description (including the radiology report, identified findings, and associated parameters), generate a highly accurate and relevant text query. The query should reflect the core findings or conditions mentioned and should be suitable for searching or clarifying information in medical databases.
>
> Instructions for the Model:
>
> Analyze the provided medical description:

Pay attention to key terms and conditions (e.g., atelectasis, absence of pleural effusion, and so on). Note that the report mentions postoperative changes, the state of the heart, and the lungs. Identify the main idea: Which problems or questions might a physician or researcher have based on these findings?

Formulate a natural-language query that:

Reflects the clinical picture described in the report (e.g., the presence of atelectasis). Can be used to further clarify the diagnosis, seek management recommendations, or find more information about the patient's condition. May include key terms such as "atelectasis," "bilateral resections," "postoperative changes," "normal heart size," etc. Ensure that the query:

Clearly states the medical issue (for example, "How to manage atelectasis following bilateral lung resections?"). Is relevant to the data from the report and accurately describes the key findings.

* Use english
* Use different sized formulations
* Use different formulations and start your query in different ways
* Output format: {'query':"}

We used Stella Zhang et al. (2024) as embedding model. This model, which is among the top performers on the MTEB leaderboard Muennighoff et al. (2022), is a specially trained variant of gte-Qwen2-1.5B-instruct Li et al. (2023) featuring a Matryoshka Representation Learning (MRL)-based Kusupati et al. (2022) method to reduce the size of the output embeddings. For contrastive learning, we used InfoNCE loss Oord et al. (2018). The fine-tuning process for Stella was performed on an NVIDIA A100 GPU, with early stopping applied to prevent overfitting. The entire fine-tuning process required no more than 5 epochs for each model configuration, ensuring efficient training.

Table 1 shows the Retrieval R-Precision metric, which measures the ability of the model to retrieve relevant items from a set of 10 candidates. These results demonstrate that fine-tuning Stella on mixed data yields the best performance, achieving a Retrieval R-Precision of 0.79 in the query-to-report task and 0.88 in the report-to-report task. Fine-tuning only on synthetic data allowed us to obtain R-Precision of 0.72 for the query-to-report task and 0.7 for the report-to-report task, significantly outperforming the baseline model (without fine-tuning). For comparison, we fine-tuned the model only on the original radiologist reports, it showed better performance than the model trained only on synthetic data and worse performance than the model trained on mixed data in both tasks.

4 Conclusion and discussion

In our study, we adapted the approach to improve embeddings using a large language model (LLM) for retrieval-based clinical decision support. Our study demonstrates that synthetic data generated by LLM can significantly enhance text embedding quality while reducing reliance on labeled datasets. We used the public available MIMIC-CXR database to generate synthetic queries. Practicing physicians were involved in generating prompts and validating the generation results.

Quality, high labeling costs, and data privacy concerns are among the key challenges hindering the development of effective AI solutions in healthcare Ahmed et al. (2023). Synthetic data offers a way to overcome these limitations, enabling the creation of tailored solutions even for niche problems. The results presented in Table 1 demonstrate that fine-tuning a model using only synthetic data can significantly enhance baseline performance. Moreover, when the model is fine-tuned on a combination of synthetic and real data, it achieves even better results, outperforming models trained exclusively on real data. We note that a relatively small LLM with 1.5B parameters was required to achieve these results.

We would like to point out that we did finetune in report-to-query mode, which allowed us to generate many queries for one patient data. This is a safer mode than query-to-report, because it does not require generating synthetic patient data. In addition, Table 1 shows that improving performance on the test dataset in the qwery-to-report task improves quality in the report-to-report task.

Synthetic data also addresses challenges related to multilingual embeddings. While high-resource languages benefit significantly, low-resource languages still require improvement due to biases in

pre-trained LLMs. Future work could focus on enhancing multilingual capabilities by incorporating more diverse pre-training corpora.

5 ACKNOWLEDGMENTS

All authors were supported by the Research Center of the Artificial Intelligence Institute of Innopolis University.

REFERENCES

Molla Imaduddin Ahmed, Brendan Spooner, John Isherwood, Mark Lane, Emma Orrock, and Ashley Dennison. A systematic review of the barriers to the implementation of artificial intelligence in healthcare. *Cureus*, October 2023. ISSN 2168-8184. doi: 10.7759/cureus.46454. URL http://dx.doi.org/10.7759/cureus.46454.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. Mimic-cxr database, 2024. URL https://physionet.org/content/mimic-cxr/2.1.0/.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning, 2022. URL https://arxiv.org/abs/2205.13147.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022. doi: 10.48550/ARXIV.2210.07316. URL https://arxiv.org/abs/2210.07316.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11897–11916. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.642. URL http://dx.doi.org/10.18653/v1/2024.acl-long.642.

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, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. Jasper and stella: distillation of sota embedding models, 2024. URL https://arxiv.org/abs/2412.19048.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific Data*, 10(1), December 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02814-8. URL http://dx.doi.org/10.1038/s41597-023-02814-8.