Generating Medical Diagnostic Scenarios with LLM-Based Reinforcement Learning Feedback: Dataset Release and Methodology

Aniruth Ananthanarayanan
University of North Texas, AniruthAnanthanarayanan@my.unt.edu

***Abstract* - Sample medical scenarios play a crucial role in training healthcare professionals by providing structured cases to develop diagnostic reasoning and clinical decision-making skills. However, access to diverse and inclusive sample diagnostic cases remains challenging due to the limited representation of specific conditions and populations in medical education materials, and existing cases are often not equitable due to a lack of representation of minority groups. In this paper, we present a new dataset of medical diagnostic scenarios generated using a combination of reinforcement learning from artificial intelligence feedback and retrieval augment generation techniques. Despite the dataset's limited size, it offers a unique resource for advancing medical education, particularly in regions with scarce training materials while also emphasizing inclusivity by incorporating a higher representation of people of color and women. Then, we discuss the data generation process, the dataset structure, and potential applications in medical training programs. This work aims to contribute to the development of accessible, high-quality, and inclusive educational tools in the medical field.**

Index Terms - Medical education, reinforcement learning from AI feedback, retrieval-augmented generation, large language models

INTRODUCTION

Accurate medical diagnosis is fundamental to effective patient care [1], yet the available educational resources are often limited both in quantity and variety [2-3]. Traditional case study scenarios in medical textbooks or reference manuals may not encompass the wide array of clinical scenarios encountered by practitioners, resulting in gaps in training. This issue is particularly glaring in low- to middle-income countries (LMICs), where access to comprehensive healthcare training material is often scarce, potentially hindering the delivery of high-quality healthcare services [4].

To address these concerns, organizations like Stanford's Clinical Mind AI lab and Laerdal have developed differentiated, innovative approaches to this problem. The Clinical Mind AI has opted for primarily software-based solutions, such as their current "Assessment of Clinical Reasoning Skills using AI-Simulated Patients: Initial Validity Evidence of the Platform Clinical Mind AI" project that aims to refine soft-skills such as patient history collection through an interactive "a targeted scoring rubric and large language models for automated assessment" [5]. Laerdal, on the other hand, has elected to build "Harvey", a cardiopulmonary simulator that helps teach "bedside assessment skills" while also promoting diversity and inclusion [6].

While these initiatives represent significant advancements in artificial intelligence-assisted medical education, they are often resource-intensive, making them less accessible to institutions in LMICs. Furthermore, existing case studies in textbooks are simplistic in nature as they often seek to explain a specific concept through a practical example. The implication of this is that these scenarios may be lacking in diversity and inclusivity, which are necessary to prepare healthcare professionals for the wide array of scenarios they may experience in practice. This limitation underscores the specific need for scalable, inclusive, and equitable training scenarios.

In this research, we present a new method of medical scenario generation powered by large language models (LLMs) in combination with reinforcement learning from artificial intelligence feedback (RLAIF) style refinement and retrieval-augmented generation (RAG) techniques [7-9]. The primary issue with simply asking an LLM to generate a medical scenario, which we refer to as naive generation, is the risk of the model generating information that is not medically accurate or relevant, either by hallucination or drift. Yang et. al. have shown that Retrieval-Augmented Generation (RAG) enables generative AI models to produce more reliable content by leveraging external knowledge sources, enhancing equity, reliability, and personalization in healthcare applications [10], while a small scale RLAIF system enables iterative improvement of a medical diagnostic scenario while reducing reliance on human input, resulting in improved scalability and efficiency.

Although the current generated dataset is very small, it contains several demonstrations of the utility of this novel generation process. The dataset includes several scenarios with people of color, women, and rare diseases, all while maintaining medical accuracy and building extensive patient profiles.

In the following sections, we will detail the generation process, describe the dataset structure, and explore potential applications in medical training programs. This initiative also aims to pave the way for more accessible and equitable healthcare worldwide.

METHODOLOGY

The dataset generation incorporated a novel process for generating medical diagnostics scenarios using small-scale reinforcement learning with AI feedback (RLAIF) and retrieval-augmented generation (RAG). The primary objective of the dataset is to make accessible clinically diverse, medically accurate, and detailed medical diagnostic scenarios that can be used to support medical training and machine learning applications. By integrating large language models (LLMs) with iterative AI-based refinement and contextual knowledge retrieval into the generation process, we aim to enhance the factual accuracy, coherence, and inclusivity of medical cases. We also aim to address concerns regarding a lack of scenarios for people of color and women by prompting models to generate data for these demographics.

Our methodology for the generation of a single case in this dataset follows a structured pipeline consisting of the following steps:

○ Vector embeddings generated from medical knowledge sources to enable RAG functionality.

○ Initial scenario generation with a high-temperature LLM.

○ Iterative small-scale reinforcement learning-based refinement via an LLM critic.

○ Structured outputting via function-calling formatting into JSON for downstream usability.

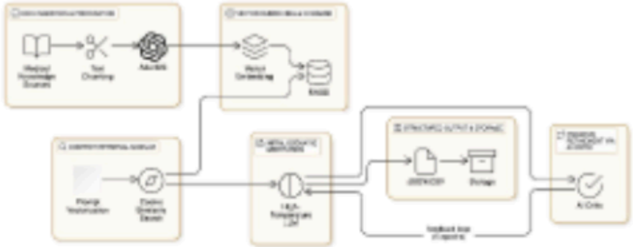

FIGURE I
SYSTEM ARCHITECTURE DIAGRAM OF PROPOSED SYSTEM

I. Vector Embedding and RAG

To enhance factual accuracy, we implemented RAG by integrating real-world medical knowledge sources [9-10]. This functionality involves two major steps: embedding generation and the retrieval process. During initialization, real-world medical resources like the *CURRENT Medical Diagnosis and Treatment* are split into chunks and converted into numerical vector representations using OpenAI's Ada 002 pre-trained sentence embedding model [11]. These sources were stored in a high-dimensional vector database with Facebook AI Similarity Search (FAISS), a scalable library for fast nearest-neighbor lookups [12]. The vector database was then cached locally.

When generating, refining, or critiquing scenarios, the prompt is converted into a dense vector representation using the same embedding model as the initialization process, and the query vector is compared against the indexed chunks' vectors. Cosine similarity is used to rank the top-k most relevant chunks [13]. Retrieved chunks are then prepended to the generation context and prompt, ensuring models are aware of relevant medical information when generating scenarios.

During development and testing, we found that there was little difference in performance in retrieving relevant chunks when comparing the Ada model and various domain-specific embeddings models such as the MedEmbed family [14].

II. Initial Scenario Generation

The first step in generating a scenario involves the creation of a baseline medical diagnostic scenario using a high-temperature large language model. This phase establishes the foundational structure of a case, and the high-temperature value helps to generate more unique cases. Initially, a topic, such as "influenza" or "skin rashes in people of color" is chosen, and relevant context is retrieved. In addition to this information, the model is prompted using the chain of thought (CoT) technique to reduce hallucinations [15]. The LLM's output is formatted into a coherent object with structured output via function calling. This coherent object contains patient history, symptoms, vital signs, diagnostic findings, diagnoses, and potential treatments. A well-documented limitation of AI-generated medical content is the underrepresentation of women and people of color in clinical scenarios. To counteract this issue, we explicitly instructed the LLM to generate cases featuring diverse patient demographics, ensuring a balanced dataset. By choosing a diverse set of topics and using a high-temperature LLM to generate initial scenarios, we curated realistic and clinically valuable cases suitable for medical education.

III. Iterative Refinement Using AI feedback

To enhance the quality, coherence, and clinical accuracy of the initially generated scenarios, we implemented an iterative refinement process with AI feedback. This technique builds on top of the core principles of reinforcement learning from human feedback (RLHF), but differentiates itself by replacing human annotators with an AI-based critic model that evaluates and refines scenarios through multiple iterations. This approach ensures scalability and detailed responses while maintaining validity.

Once an initial diagnostic scenario is generated, it undergoes an automated quality assessment by a specialized RAG LLM critic prompted to identify errors and inconsistencies through CoT prompting. The critic is instructed to assess the following metrics:

- Medical Accuracy: Does the case align with known medical knowledge?
- Coherence: Is the case logically structured?
- Completeness: Does the scenario include all necessary components, and is there any way to make it more detailed?
- Bias and Inclusivity: Does the scenario fairly represent diverse patient demographics?

The AI critic identifies potential areas for improvement, and the scenario, along with feedback, is sent back for revision.

IV. Dataset Structuring and Storage

Once scenarios had gone through 5 epochs of refinement, the structured output was converted to a JSON object and stored in a .json file to maximize usability for downstream applications.

V. Limitations and Implementation Notes

The existing dataset is only 100 scenarios large. Further efforts will be focused on extending the size of the dataset and increasing the number of refinement iterations to improve the quality and detail of the generated examples. The pipeline is currently built in Python and is accessible through a command line interface. The model used to generate the dataset was GPT-3.5-Turbo, and the system was orchestrated with Langchain. External documents were split into 1000-character chunks with 600-character overlap, using the OpenAI Ada 002 embedding model.

Results

In this section, we present preliminary results comparing three approaches to generating diagnostic scenarios: the "naive" method, a simplistic RAG with no refinement, and our approach, which we refer to as RAG + Critic. Two primary metrics were used for evaluation: medical accuracy and detail. These metrics were calculated by manually comparing a representative sample of the generated cases and taking the average of the ratings for each metric for each of the methods.

TABLE I
Preliminary Results

Method	Medical Accuracy (0-10)	Detail (0-10
Naive	9.59	5.59
RAG	10	5.56
RAG + Critic	10	5.78

While these initial results show promising improvements in medical accuracy, and modest gains in detail with the RAG + Critic method, future work will incorporate more comprehensive evaluations, with planned enhancements including expert evaluations to provide further insights into system performance and to refine the methodology specifically for medical education (i.e. through additional metrics such as "educational value").

Discussion

The proposed dataset and generation methodology address several key challenges in medical education, particularly regarding diversity, accessibility, and accuracy in diagnostic training materials. By leveraging retrieval-augmented generation and reinforcement learning from AI feedback,, we introduce a novel approach that enhances the factual accuracy, coherence, and inclusivity of medical scenarios. However, beyond its direction application in medical education, this methodology presents opportunities for broader implementations in other disciplines, pre-college stem initiatives, and use in LMICs.

I. Comparison to Existing Methods

Traditionally, diagnostic scenarios in medical education are crafted by domain experts with extensive clinical experience and a deep understanding of patient care. However, expert-generated content is limited in terms of scalability, subjectivity, consistency, and resources. Creating a single expert scenario is labor-intensive, making it challenging to produce large, diverse datasets needed for comprehensive training. In addition to this, while experts provide valuable insights, hand-written cases risk containing human biases, while variations in expertise may lead to inconsistencies across cases. In contrast, our automated method aims to bridge these gaps by delivering scalable, consistent, and rapid generation of diagnostic scenarios. Although our method might not yet capture all the subtle nuances of expert-crafted cases, it offers a reproducible, efficient, and scalable alternative. Moreover, model-specific fine-tuning and the inclusion of direct expert feedback into the AI refinement process in future iterations show promise in further aligning outputs with expert insights.

Additionally, our approach distinguishes itself by integrating domain-specific context enrichment with an iterative AI feedback loop. Unlike the traditional "naive" generation approach, which often lacks sufficient contextual grounding, combining RAG with a critic enables an enriched prompt with relevant medical context and effectively contains an example of what a good scenario is to be efficiently augmented, resulting in improved factual accuracy and a stronger foundational structure for each scenario. Most importantly, however, is the method's advantage over a simplistic RAG approach. One of basic RAG's pitfalls is its inability to retain information on relationships. By integrating a RAG critic that evaluates scenarios against diverse real-world medical snippets, our approach effectively captures these critical relationships.

II. Applications in Other Disciplines

The combination of RAG and refinement techniques detailed in this research offers a scalable and adaptable approach that extends beyond medical education into other STEM fields. The ability to iteratively refine AI-generated content using domain-specific knowledge retrieval and small-scale reinforcement learning presents opportunities for enhancing content generation in disciplines requiring accurate, context-aware problem-solving scenarios.

In engineering education, AI-driven case studies can be generated and refined to produce realistic simulations of structural failures, thermodynamic analyses, or circuit design challenges. Contrary to traditional methods, simulation-based methods could provide an opportunity for students to engage with real-world conditions and reason about new problem-solving approaches in controlled conditions [16]. By retrieving relevant engineering principles and reinforcing the AI-generated content through iterative feedback loops, students can engage with dynamically updated problem sets that evolve based on expert-reviewed constraints and industry standards. This ensures the learners are exposed to increasingly complex scenarios tailored to real-world engineering applications.

In computer science education, this system offers benefits in generating and refining programming exercises, cybersecurity attack-defense scenarios, and helping students learn algorithmic principles dynamically. AI generation also stands out in its ability to cater to students' learning styles. Human input specifying topics and difficulty in combination with problem generation with this framework can help create more optimal learning trajectories for students engaging with advanced computing concepts. This technique can also be used to prepare for standardized tests in general. By grounding sample problems in existing exams, students can more effectively adapt to the difficulty level of the test, which is often not exact in third-party sample exams [17].

The adaptability of this methodology highlights its potential as a transformative tool in interdisciplinary STEM education. Using this framework, educators can enhance engagement and improve learning outcomes, bridging the gap between theoretical knowledge and real-world application.

III. Pre-College Initiatives and Outreach Programs

AI-driven educational tools have the potential to enhance pre-college initiatives and outreach programs by making complex concepts more accessible and engaging. Less complex versions of the generated scenarios can help introduce high school students to problem-solving techniques used in medicine, engineering, and data science. Interactive learning experiences, such as virtual lab simulations, can provide students with hands-on experience, helping them refine technical skills in a meaningful manner. Additionally, the integration of artificial intelligence-based teaching techniques can help bridge educational gaps in underserved communities, ensuring that students with unique perspectives from diverse backgrounds have exposure to advanced STEM topics before entering higher education.

IV. Potential Impacts in LMICs

The implementation of AI-generated medical training datasets has profound implications for LMICs, where access to quality educational materials is often limited. By leveraging AI to create a scalable and cost-efficient system to generate diverse, medically accurate diagnostic cases, this methodology can provide low-cost resources for healthcare training programs in LMICs. Furthermore, AI-driven medical education tools can supplement traditional learning methods through simulation-based learning, which has been shown to improve patient outcomes [18]. This becomes extremely valuable in areas with a shortage of trained instructors, ensuring that students receive high-quality instruction regardless of geographical or financial constraints. Future research could explore partnerships with global health organizations to build a larger infrastructure to support the distribution of AI-generated medical scenarios to healthcare training institutions worldwide.

V. Synthetic Patient Data Generation for AI/ML Applications

Ensuring patient confidentiality is a critical aspect of medical data usage. By employing synthetic data generation techniques, end users can circumvent risks associated with using real data. By incorporating a more in-depth corpus of medical knowledge, this methodology can create realistic medical diagnostic scenarios without compromising patient privacy, without compromising patient privacy. Synthetic data generation can also make AI/ML models trained on them more ethically viable through the controlled inclusion of more diverse demographic characteristics, helping them become more applicable across demographics. This approach enhances the usability of synthetic data while aligning with ethical standards and simplifying regulatory requirements for medical data protection.

VI. Multimodal Data Generation & Physical Devices

An important extension of this methodology lies in the potential for generating multimodal data, combining the text-based diagnostic scenarios with other forms of data, such as medical imaging or patient audio recordings. In medical education, case studies are often enriched by visual data, such as X-rays, MRIs, or ultrasound images, which help future healthcare professionals develop reasoning skills more comprehensively. Integrating these modalities can extend the utility of the dataset and provide a more holistic learning experience, where cases are paired with relevant images, diagnostic graphs, or audio of patient symptoms (e.g., breathing sounds for respiratory diseases or X-ray for a broken wrist).

Multimodal data generation requires an advanced pipeline that requires models to generate complex text scenarios, determine potential visual or audio data sources, and accurately generate plausible images or sounds. This requires high-level relational reasoning and a higher-level understanding than can be achieved through RAG. A proposed framework to potentially do this would most likely integrate a state-of-the-art reasoning model functioning as the head of a multi-agent system, similar to that of the virtual lab developed by Swanson et. al. [19].

Multimodal data can enhance the richness and depth of training materials, ensuring that medical professionals are exposed to diverse learning experiences through pure software. However, another alternative approach to creating a more holistic educational resource is the development of a physical device to run and "display" the generated scenarios from this methodology. Similar to Harvey, the framework in this paper, combined with multimodal audio generation, could enrich the quality of education by requiring students to refine their data collection skills along with their diagnostic skills [6].

CONCLUSION

The dataset and methodology in this research have broad implications beyond medical education, offering opportunities to both transform STEM learning and evolve into a more holistic method of educating students to handle real-world scenarios. By leveraging AI-driven content generation, educational institutions can create more inclusive, adaptive, and scalable learning resources that address gaps in accessibility, diversity, and equity. Future efforts should focus on expanding the dataset, refining AI-generated educational tools, and assessing their impact on student learning and professional outcomes. As AI continues to revolutionize STEM education, ensuring the development of high-quality, inclusive, and ethical training materials remains a critical priority.

DATASET AVAILABILITY

The dataset generated in this research is available for academic and educational use. Those interested in accessing the dataset can reach out to me directly for inquiries. Additionally, the dataset will be made available on GitHub in the future, where updates, expanded case studies, and refinements based on feedback from medical professionals and educators will be provided.

REFERENCES

[1] Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine, & The National Academies of Sciences, Engineering, and Medicine. (2015). *Improving diagnosis in health care*. National Academies Press. https://www.ncbi.nlm.nih.gov/books/NBK338593/

[2] Densen, P. (2011). Challenges and opportunities facing medical education. *Transactions of the American Clinical and Climatological Association, 122*, 48–58.

[3] Awosogba, T., Betancourt, J. R., Conyers, et. al. (2013). Prioritizing health disparities in medical education to improve care. *Annals of the New York Academy of Sciences, 1287*, 17–30. https://doi.org/10.1111/nyas.12117

[4] Phelan, H., Yates, V., & Lillie, E. (2022). Challenges in healthcare delivery in low- and middle-income countries. *Anaesthesia and Intensive Care Medicine, 23*(8), 501–504. https://doi.org/10.1016/j.mpaic.2022.05.004

[5] Stanford Clinical Mind AI. (n.d.). *Research*. Stanford University. https://clinicalmindai.stanford.edu/research-0

[6] Laerdal Medical. (n.d.). *Harvey: The cardiopulmonary patient simulator*. https://laerdal.com/us/harvey/

[7] Naveed, H., Khan, A. U., Qiu, S., et. al. (2024). *A comprehensive overview of large language models*. arXiv [Cs.CL]. http://arxiv.org/abs/2307.06435

[8] Lee, H., Phatale, S., Mansoor, H., et. al. (2024). *RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback*. arXiv [Cs.CL]. http://arxiv.org/abs/2309.00267

[9] Lewis, P., Perez, E., Piktus, A., Petroni, et. al. (2021). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. arXiv [Cs.CL]. http://arxiv.org/abs/2005.11401

[10] Yang, R., Ning, Y., Keppo, E. *et. al.* Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Syst.* 2, 2 (2025). https://doi.org/10.1038/s44401-024-00004-1

[11] Neelakantan, A., Xu, T., Puri, R., et. al. (2022). *Text and code embeddings by contrastive pre-training*. arXiv [Cs.CL]. http://arxiv.org/abs/2201.10005

[12] Douze, M., Guzhva, A., Deng, C., et. al. (2024). *The Faiss library*. arXiv [Cs.LG]. http://arxiv.org/abs/2401.08281

[13] Steck, H., Ekanadham, C., & Kallus, N. (2024, May). *Is cosine-similarity of embeddings really about similarity? Companion Proceedings of the ACM Web Conference 2024*, 887–890. https://doi.org/10.1145/3589335.3651526

[14] *MedEmbed: Fine-tuned embedding models for medical / clinical IR*. Hugging Face – The AI community building the future. (n.d.). https://huggingface.co/blog/abhinand/medembed-finetuned-embedding-models-for-medical-ir

[15] Barkley, L., & van der Merwe, B. (2024). Investigating the Role of Prompting and External Tools in Hallucination Rates of Large Language Models. *arXiv [Cs.CL]*. Retrieved from http://arxiv.org/abs/2410.19385

[16] Balamuralithara, B. and Woods, P.C. (2009), Virtual laboratories in engineering education: The simulation lab and remote lab. Comput. Appl. Eng. Educ., 17: 108-118. https://doi.org/10.1002/cae.20186

[17] Shah, V., Yu, D., Lyu, K., et. al. (2025). AI-Assisted Generation of Difficult Math Questions. *arXiv [Cs.AI]*. Retrieved from http://arxiv.org/abs/2407.21009

[18] Elendu, C., Amaechi, D. C., Okatta, A. U,. et. al.. (2024). The impact of simulation-based training in medical education: A review. *Medicine*, *103*(27), e38813. https://doi.org/10.1097/MD.0000000000038813

[19] Swanson, K., Wu, W., Bulaong, N. L., et. al. (2024). The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation. *bioRxiv*. doi:10.1101/2024.11.11.623004