Annals of Emerging Technologies in Computing (AETiC)
Vol. 9, No. 5, 2025

Research Article

Clinical Text Augmentation and Generation Using RAG for Large Language Models

Nasreen Fathima and Kavitha Ganesh*

B. S. Abdur Rahman Crescent Institute of Science & Technology, India
nasreenmansoor2001@gmail.com; gkavitha.78@gmail.com
*Correspondence: gkavitha.78@gmail.com

Received: 19th January 2025; Accepted: 27 July 2025; Published: 25 October 2025

Abstract: Large Language Models (LLM) are becoming more essential in clinical text generation, where use of synthetic medical data is environmentally accurate and applicable for real-world healthcare applications. Existing LLMs often lack in specialized optimization and clarity, leading to incorrect outputs. These restrictions can make their references unreliable, particularly for sensitive clinical data. To overcome these problems, this research work suggests integrating generative adversarial networks with LLM to improve clinical data accuracy and reduce hallucinations. LLMs like LLaMA, BERT and GPT are broadly used in clinical settings for tasks such as summarizing patient notes and answering medical queries. Generative Adversarial Networks (GANs) are used to generate realistic synthetic clinical data, aiding privacy and data augmentation. The LDA model is added with GAN to identify the underlying topics in clinical documents, ensuring the synthetic text is coherent and thematically relevant. The use of Retrieval Augmented Generation (RAG) dynamically retrieves current medical knowledge and provides grounding responses with real-time evidence and minimizes outdated information. The first phase focuses on generating and validating synthetic clinical data using GANs and LDA to ensure high quality and domain alignment; the second phase focus on user interaction, where RAG retrieves relevant information in real time to answer queries, and an interactive interface enables seamless engagement and feedback. Continuous evaluation of NLP metrics demonstrates that the proposed Clinical Augmentation Generation and Retrieval Augmented Generation (CAG-RAG) framework outperforms the existing DALL-M approach in generating synthetic clinical text. For diagnosis-related data, the proposed CAG-RAG method achieves improvements of 15.7% in BLEU, 17% in ROUGE-1, and 17% in ROUGE-L scores. For medication-related data, the improvements were 20.8% in BLEU, 17.1% in ROUGE-1, and 17.25% in ROUGE-L. These results highlight the reliability, adaptability, and contextual accuracy for clinical applications.

Keywords: ***Large Language Models; Retrieval Modules; Synthetic Clinical Data; Data Augmentation; Synthetic Text; NLP Metrics; Generative Adversarial Network; Retrieval Augmented Generation***

1. Introduction

Incorporating artificial intelligence in healthcare has opened a door to novel applications of clinical decision-making and medical research. Yet for AI models to work, it needs to provide responses that are both contingently correct and clinically useful. While synthetic data generation alleviates issues such as data thinness and privacy, the calibration of information is necessary to make it coherent and factually precise. Large Language Models play a vital role in healthcare sectors to generate clinical text data through synthetic data generation to aid the medical practitioners in diagnosis, treatment and medical research [1].

Large Language Models based validation and refinement are required for generating texts that are more accurate [2-3]. Transformer based models such as LLaMA, BERT, and GPT are used in applications involving sequential data to evaluate the semantic correctness of the generated text. However, the LLMs used for generating clinical responses should remain factual and up-to-date and hence retrieval based

Nasreen Fathima and Kavitha Ganesh, "Clinical Text Augmentation and Generation Using RAG for Large Language Models", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 61-70, Vol. 9, No. 5, 25 October 2025, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2025.05.005, Available: http://aetic.theiaer.org/archive/v9/v9n5/p5.html.

augmentation techniques are employed. Retrieval Augmented Generation, dynamically fetches the relevant documents for generating a relevant response. Compared to traditional methods that are dependent on pre-trained models, RAG retrieves clinical data in real time for making quick responses to the documents and contextually more relevant. It also eliminates the possibility of AI-generated hallucinations and enhances the reliability of the system in clinical use. Evaluation metrics such as BLEU score is used to access the quality of machine generated text, ROUGE 1 score is used to assess the quality of text summarization and ROUGE L scores measures the length of longest subsequence of words to measure text quality [4-5].

Presently, the LLMs are trained on clinical data for improving diagnosis and suggestive responses in healthcare. Due to the data scarcity in clinical data, synthetic data that structures and mimics the real clinical data is generated by Generative Adversarial Networks [6]. LLMs with synthetic data finds its application in disease classification and clinical decision making for medical practitioners. Integrating Latent Dirichlet Allocation (LDA) for topic modeling in clinical text, improves the diversity and quality of synthetic data by capturing latent patterns within the generated synthetic data [7-8]. Further the LLMs aids context aware feature augmentation to generate additional synthetic features thereby providing context-specific responses to user queries. Further, by leveraging the information from the generated synthetic data and external knowledge sources with zero shot learning, the proposed CAG-RAG model provides intelligent query recommendations and context-specific clinical responses. Hence, by integrating the generative adversarial networks with retrieval augmented generation delivers a robust, privacy-preserving solution for clinical information retrieval and decision support in health care systems.

2. Related Work

Clinical data is essential in healthcare for medication, treatment, diagnosis, and investigation but it is being controlled by factors such as data shortage and privacy laws. Synthetic data generation solves this problem by producing false data that replicates the statistical characteristics of true data without violating privacy. Generative Adversarial Networks employ a generator and discriminator to generate realistic synthetic data and have proven effective in medical imaging as well as clinical text generation. Latent Dirichlet Allocation (LDA), a topic modeling technique, facilitates the extraction of underlying topics from text [9]. Coupling LDA with GANs improves the diversity and contextually relevant nature of the synthetic clinical data.

Hsieh *et al.* [10] proposed the idea of integrating imaging and clinical data, which lacks patient-specific context. Deep learning models and conventional methods suffer data integrity and misinterpretation of rare diseases. To mitigate this, DALL-M framework was introduced that creates context-aware synthesized clinical features from X-rays and patient reports. DALL-M works in three stages namely, clinical context repository, expert query generation, and context-aware augmentation resulting in better dataset richness and model performance.

Latif and Kim [11] discussed the data scarcity challenge in deep learning based clinical healthcare systems. Classical augmentation techniques such as back-translation tend to lack contextual fidelity. The authors investigate ChatGPT-based augmentation for improving the CHARDAT dataset, producing contextually aligned but semantically different clinical instances. In combination with ChatGPT, other techniques such as EDA and AEDA were evaluated, where models such as BART registered better ROUGE scores. The research shows the ability of LLMs to generate high-quality synthetic data, which greatly enhances model performance in clinical settings.

Kim *et al.* [12] suggested a semi-supervised image captioning framework to minimize dependence on labor-intensive paired datasets. The proposed approach uses unpaired unimodal data, images or captions obtained independently by employing adversarial learning to infer pseudo-labels and learn their joint distribution. The framework demonstrates robust generalization, for out-of-task or web-sourced data, and makes consistent gains on benchmarks such as COCO dataset. Hence, semi-supervised learning is able to effectively fill the gap between paired and unpaired data, enhancing image captioning performance while reducing the cost of annotations.

Ghanadian *et al.* [13] outlined the issues of ethical concerns and data scarcity that prevents the development of effective ML models in constrained studies. Existing NLP techniques, including BERT models have low performance due to overfitting and data scarcity. Generative AI models such as ChatGPT,

Flan-T5, and Llama are employed to create synthetic data from psychological and social states. Integrating about 30% of real data with synthetic data has achieved an F1-score of 0.88, showing improved model performance and diversity. The study outlines the application of synthetic data in sensitive environments and suggests cultural, linguistic, and multimodal adaptation as future research directions. Saman Motamed *et al.* present IAGAN, a cutting-edge GAN-based approach to overcome data scarcity in medical imaging, specifically for diseases like pneumonia and COVID-19[14]. Compared to traditional augmentation and GANs like DCGAN, IAGAN uses convolutional and attention layers to generate class-specific synthetic images like chest X-rays from partially labeled samples. Empirical results achieve AUC gains of 2–3% and sensitivity/specificity gains of 1–3%. IAGAN is particularly effective in low-data, high-response scenarios, but challenges remain to overcome multi-centric variability in datasets like COVIDx.

Jelodar *et al.* [15] presents a comprehensive overview of Latent Dirichlet Allocation, one of the most significant topic modeling methods widely used in extracting latent patterns from unstructured text. The paper surveys LDA-based studies between 2003 and 2016, focusing on its usage across diverse fields including medical research, political science, and software engineering. While offering valuable insights into document organization by modeling topics as probabilistic word distributions, LDA suffers from semantic understanding, high computational complexity, parameter tuning, and interpretability. These limitations have prompted the development of more advanced topic modeling methods and tools to address domain-specific needs.

Arora and Arora [16] discussed on the potential revolution in medical domain using Generative Adversarial Networks (GANs) to generate synthetic data while ensuring patient confidentiality, solving ethical problems of data deficiency, and reducing bias. GANs allow secure data transfer, enabling realistic clinical cases to be generated for medical education. While this is a promising solution, there exists ethical issues of misuse, false content generation, and attribution to AI which needs to be addressed. With appropriate deployment, GANs have the potential to transform clinical research, education, and patient care.

Biswas *et al.* [17] proposed the use of Generative Adversarial Networks as a new medical data augmentation technology, especially when real data is scarce or ethically unavailable. GANs iteratively train a discriminator and generator to produce synthetic data that is similar to real samples. It includes data augmentation, copying generation, and domain adaptation. GANs have vast potential to combat data scarcity, but their rigorous evaluation is still required to determine the validity and clinical quality of the generated images.

Imtiaz *et al.* [18] propose a GAN-based framework with differential privacy to address the issue of secure data sharing in smart healthcare systems. By learning from the data distribution, the framework generates privacy-preserving synthetic datasets with preserved statistical characteristics of the original data. On a real-world Fitbit dataset, the framework is able to strike a balance between data utility and privacy, promoting safe and open data sharing for research and industry applications.

Kumichev *et al.* [19] introduces MedSyn, to generate high quality synthetic clinical notes. With GPT-4 and fine-tuned LLaMA models, MedSyn improves the accuracy of ICD code classification by 17.8% for hard to classify codes compared to real data. MedSyn also generates the largest open source dataset of synthetic Russian clinical notes with 219 ICD-10 codes expanding disease classes for clinical decision support systems. The system is beneficial for low-resource languages and tasks like NER tagging and ICD coding for publicly available models and datasets.

Existing research faces challenges in using clinical data for machine learning and AI tasks. Most available clinical datasets are small and have class imbalance, which affects model generalization and accuracy. The use of real patient data poses serious privacy threats, potentially violating sensitive health information. In response to these challenges, researchers have investigated synthetic data generation; yet, this process fails to capture the complete nuance of actual clinical situations. Furthermore, fine-tuning large-scale models on clinical data is a computationally intensive process that consumes a lot of resources, and

hence it is expensive and less practical for universal use. These constraints emphasize the necessity for better, privacy-protecting, and computationally inexpensive solutions for clinical data modeling.

3. Proposed Architecture

This work proposes a scalable method of creating high-quality synthetic clinical data through the combination of GAN and LDA. The major problems in clinical data analysis include data inadequacy, class inequality, and data privacy. To combat the above issues, it is proposed to generate synthetic data equivalent to real clinical data while preserving domain-specific information. Generative Adversarial Networks create synthetic clinical data by implementing a generator and a discriminator. The generator generates data samples by learning intrinsic patterns and discriminator verifies the genuineness of the generated data. Through the adversarial feedback loop, the level of factual representation in generated data is improved, and such data can be applied in clinical environments where heterogeneous and large sets of data are not usually accessible due to privacy limitations.

To boost the fidelity and range of synthetic data, gaussian noise is added to the original data. This simulates the natural variability of real clinical environments, resulting in more generalized and robust data that is not overfitted to the original data. The model also incorporates the usage of LDA for topic modelling and feature engineering. LDA reveals hidden topics in clinical text, i.e., treatments, symptoms, or medical conditions, that authenticates synthetic data adheres to domain-specific medical expertise. The incorporation of LDA enables synthetic data generation for underrepresented classes to balance the dataset and enhance the performance of the model for rare conditions. Also, the privacy-preserving nature of the framework ensures that it is HIPAA compliant. The synthetic data produced contains no personally identifiable information and is therefore safe for medical and allied research. Additionally, the system employs a Retrieval Augmented Generation (RAG) process in combination with a pre-trained LLM to generate intellectual clinical responses utilizing synthetic data and external medical knowledge bases. A zero-shot learning process allows the system to handle new clinical queries without requiring extensive fine-tuning, and a query verification process guarantees the factual accuracy and medical pertinence of generated responses.

The proposed Clinical Augmentation Generation - Retrieval Augmented Generation (CAG-RAG) framework has been designed as a combined app[1]roach to offer accessible clinical data through augmentation and a decision support solution for the application of synthetic clinical data in medical research.

3.1. CAG-RAG Architecture

The system architecture of CAG-RAG for preserving data privacy and clinical decision making is depicted in Figure 1. The framework uses a clinical dataset retrieved from publicly accessible source, Kaggle repository. Raw data is pre-processed using gaussian noise vector for data consistency. The pre-processed data is applied to LDA generator to determine latent semantic patterns like disease, medication, symptoms, and treatment. These latent topics are combined with gaussian noise to produce structured inputs to a Generative Adversarial Network.

GAN comprises a generator producing synthetic clinical records and a discriminator predicting the likelihood of whether such records are authentic or not, compared to real data. The generator keeps improving its ability to produce high-quality, synthetic clinical data through adversarial training mechanism.

This process addresses problems such as data insufficiency, class imbalance, and patient privacy concerns, producing a high-volume, well-balanced, and anonymized synthetic dataset suitable for healthcare machine learning tasks. The second step of the CAG-RAG framework is an intelligent question-answering system with retrieval augmentation. Using the synthetic dataset created in the first step, clinical documents are represented as dense vector embeddings and stored in retrieval-conducive format. Upon user input of a natural language question, a dense retriever queries the synthetic corpus to find the most semantically relevant documents. These documents, along with the user question, are then passed through

[1] https://www.kaggle.com/datasets/rohitphalke1/clinicaldata

a pre-trained Large Language Model (LLM), which produces a context-aware, factually grounded response. By conditioning on both the query and the retrieved content, the system guarantees high accuracy, contextual relevance, and clinical validity in the response.

The RAG component ensure that the responses are maintained by evidence by regaining the most appropriate clinical publications and incorporating them with the user review. Using its massive medical knowledge, the pre-trained LLM, such as GPT, LLaMA, or BioBERT, understands the question and produces relevant responses. Without further fine-tuning, Zero-Shot Learning allows the system to address original and unexpected clinical queries.

Even for unusual diseases or newly evolving medical circumstances, the model can produce consistent results by exploiting semantic, intellectual, and prior data. This feature lowers reliance on large labelled datasets, which are commonly hard to obtain in the medical field. Additionally, it pledges that the model may adapt to updating medical knowledge without lacking repeated retraining. The query verification layer confirms that the generated responses are precisely correct and clinically harmless. To preserve integrity, it validates that outputs match the medical indication that was recovered, makes sure that medical terminology is precise, and eliminates responses that are inconsistent or lack supporting data.

The CAG-RAG model provides an end-to-end, privacy-protecting pipeline that not only augments clinical datasets with realistic synthetic data but also enables smart and secure access to clinical knowledge and retrieval techniques.

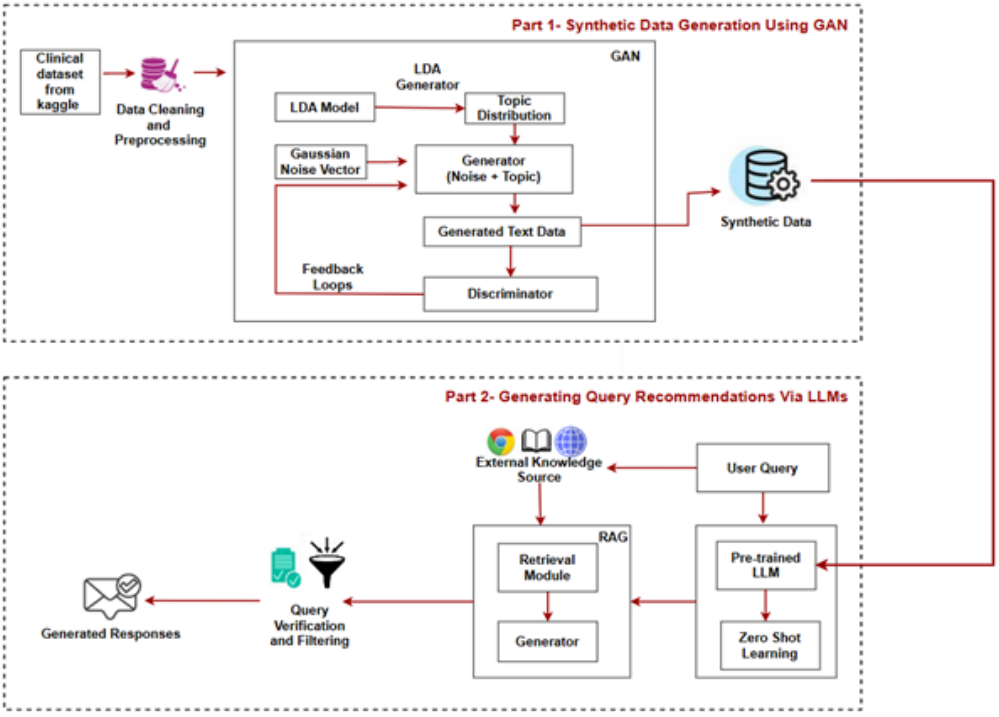

Figure 1. Overview of CAG-RAG Architecture

3.2. Data Collection and Pre-processing of CAG-RAG

The clinical text data set has been obtained from an authorized source such as Kaggle repository. The raw data is pre-processed using gaussian noise vector to ensure consistency and accuracy. The cleaned data is then tokenized for lemmatization and stemming to remove stop words for preserving semantic meaning with contextual understanding. The processed structured data is divided into training and testing subsets. The work incorporates topic modeling using Latent Dirichlet Allocation (LDA) and generate synthetic data using Generative Adversarial Networks.

3.3. LDA Topic Modelling

Latent Dirichlet Allocation (LDA) is a NLP technique used to find topics in a collection of text documents which is generated by a statistical process. Here each topic is defined by a set of words, and the

probability of a word appearing in a topic. The step wise procedure for LDA topic modelling is summarized below:

Step 1: The raw dataset is Pre-processed and cleaned
preprocessing_doc = preprocessing(raw_docs)
Step 2: LDA topic modelling trains the LDA model to learn topics and word distribution.
lda_model = LdaModel (corpus, num_topics=K, id2word=dictionary)
Step 3**:** Convert the corpus into a Bag-of-Words (BoW) format using the dictionary.
dict = Dictionary (df ['processed_clinical note'])
clinical corpus = [dict.doc2bow(texts) for text in
df ['processed_clinical notes']]
Step 4: Save topic distributions and stock topic mixtures for each document.
top_vect = [lda_model.get_document_top(doc) in corpus]

3.4. GAN-based Synthetic Data Generation

A Generative Adversarial Network is used to generate synthetic clinical records that mimic real-world patterns of data. The generator generates new samples from random inputs, and the discriminator assesses their authenticity for model refinement. The process of synthetic data generation using generator and discriminator is illustrated below:

Step 1: The generator generates synthetic clinical data by taking random noise as input.
generator = EnhancedGenerator (latent_dim, noise_dim, topic_dim)
Step 2: The discriminator receives both real clinical data and fake data from the generator, and tries to distinguish between them.
discriminator = EnhancedDiscriminator (topic_dim)
Step 3: The generator and discriminator are trained together. The generator aims to produce more realistic data, while the discriminator improves at detecting fake data.
Step 4: Based on the discriminator's feedback, the generator improves its ability to create more realistic clinical data.
for epoch in range (n_epochs):
fake_data = generator (latent_vec, noise)
d_loss = discriminator_loss (real_data, fake_data)
g_loss = generator_loss (fake_data)
Step 5: After several iterations, the generator produces synthetic data that is more realistic as real clinical data.

3.5. LLM Validation and Refinement

A transformer language model processes input queries and supporting information to produce human-like answers [20]. It works in an autoregressive fashion, predicting the next word based on previous input and learned patterns of language. The LLM validation process is depicted below:

Step 1: Load pre-trained transformer models (GPT, BERT, LLaMA) fine-tuned for clinical domain understanding.
Step 2: Input synthetic clinical text generated by the GAN into the validation pipeline.
Step 3: Detect errors in grammar, medical terminology, and ambiguous statements using the transformer models.
Step 4: Automatically correct grammatical issues and refine medical terms for clarity and consistency.
Step 5: Verify the corrected text for semantic alignment with real-world clinical data.

3.6. Retrieval-Augmented Generation(RAG) For Query Processing

The RAG framework incorporates retrieval mechanism to generate text by first retrieving the most relevant documents by an encoded user query. These retrieved entries are then fused with the query and passed to a language model for context-aware response generation. The RAG based query generation along with its generated text with reference to an external knowledge source is depicted in Figure 2. The step-by-step process of query processing mechanism is shown below:

Step 1: Query Encoding Encodes the user clinical query using a transformer encoder.
query_embed = encode (user_clinical query)
Step 2: Knowledge Retrieval Retrieve top-k docs from synthetic data.
retrieved_doc = vector_db. search (query_embed, top_k=5)
Step 3: Context Construction Concatenate the user query with the retrieved context.
context_input = concat (user_query, retrieved_docs)

Step 4: Response Generation Generate output using the LLM conditioned on the context.
response = generative_model. generate (context_input)

```
Based on the provided information, here's a breakdown of the situation and suggested treatment considerations:

**Patient Presentation:**

*   **Symptoms:** Headache, shortness of breath, fatigue (consistent with bronchitis)
*   **Diagnosis:** Bronchitis
*   **Medication:** Amoxicillin (an antibiotic)

**Treatment Considerations & Suggestions:**

1.  **Antibiotic Evaluation:**
    *   The notes state that the patient is currently taking Amoxicillin. As the patient has bronchitis, this is a c
    *   It is important to assess how the patient is doing on Amoxicillin. Is their condition improving? If not, the

2.  **Symptom Management:**
    *   **Shortness of Breath:**
        *   Assess the severity of the shortness of breath. If it is severe, getting more oxygen from a doctor might
        *   If the patient is a smoker, advise to quit.
    *   **Headache and Fatigue:** These are common symptoms of illness. Encourage rest, hydration, and over-the-cour
    *   **Cough Management:** The doctor might have given the patient cough medicine to manage the cough.

3.  **Monitoring:**
    *   **Follow-up:** Instruct the patient to follow up with the doctor. If there are any worsening symptoms.
    *   **Observe for Complications:** Watch out for worsening symptoms like:
        *   High fever
        *   Severe chest pain
        *   Worsening shortness of breath
        *   Confusion
    *   **Improvement:** Watch out for the patient improving within a week or two. If not, the doctor should evaluat

**Important Disclaimer:** I am an AI Chatbot and not a medical professional. This information is for educational pur
```

Figure 2. RAG Based Query Generation using an external knowledge source

3.7. Query Generation (Automated User Input Simulation)

Queries are assembled by extracting insignificant terms from topic vectors or clinical texts using keyword ranking techniques. These terms are structured into natural language questions to mimic user interactions and prompt the response pipeline. The process of query generation is shown below and the response for the user query is presented in Figure 3.

Step 1: Remove Keywords and then Select key terms using topic vectors or TF-IDF.
keyword = ext_keyword(docs)
Step 2: Rank and Filter the Selected top-n relevant terms for the query.
top_keyword = ranking_keyword (keyword, n=5)
Step 3: Formulate Query Construct a natural language question using selected keywords.
query = formatting_query (top_keyword)

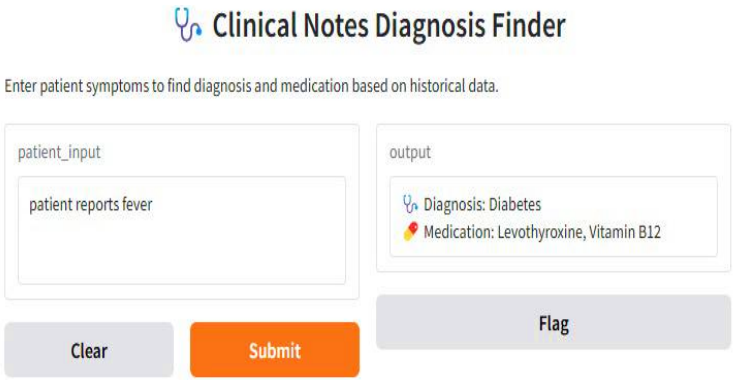

Figure 3. Query Generation

4. Results and Discussion

4.1. Experimental Setup

The CAG-RAG architecture was implemented and tested on a clinical dataset downloaded from Kaggle repository, comprised of unstructured medical notes. The experiment was performed on Google Colab, utilizing its cloud-based GPU infrastructure to enable efficient training and execution. The runtime was set to utilize an NVIDIA T4 GPU, which greatly speed up the training process in comparison to CPU-

based runs. The implementation used Python as the primary programming language, communicating with libraries like TensorFlow, PyTorch, Gensim, and Hugging Face transformers for deep learning, topic modelling, and text generation.

The synthetic data generation process comprises of data cleaning and preprocessing the clinical notes, followed by the applying LDA to derive topic distributions. The topic vectors were mixed with gaussian noise inputs and fed into a GAN-based generator to generate synthetic text, while a discriminator network provided feedback to improve the output. The training was performed for 1000 epochs with a batch size of 64, utilizing the Adam optimizer. The generated synthetic dataset, comprising more than 11,000 records, was utilized to train a Retrieval-Augmented Generation (RAG) module along with a pre-trained LLM. For each user query, the context relevant to the query was retrieved through FAISS-based vector search and fed into the LLM to generate clinically relevant responses.

4.2. Interpretation of Results

The metrics employed to measure the performance of the proposed CAG-RAG approach in synthesizing clinical text include BLEU, ROUGE-1, and ROUGE-L for measuring the quality, fluency, and contextual appropriateness of generated text against reference standards.

In the diagnosis-related clinical data case, the CAG-RAG system shows a 15.7% BLEU improvement, reflecting a considerable n-gram overlap increase between the output and reference data. This implies enhanced syntactic correctness similar as real-world clinical diagnosis note structure. In addition, the 17% improvements in ROUGE-1 score and 17% in ROUGE-L score indicate the system's capacity to measure both unigram-level similarity and longer sequence coherence, respectively, thus improving semantic relevance and contextual integrity.

For medication-related information, the gains in performance are even more significant. The model registered a 20.8% gain in BLEU score, which indicates an even greater syntactic similarity and vocabulary usage with respect to the reference texts. Similarly, 17.1% and 17.3% gains in ROUGE-1 and ROUGE-L scores respectively indicate the model's strong capability to ensure consistency and meaningfulness in medication-related sentence generation. Table 1 shows the improvements in the evaluation metrics namely BLEU, ROUGE and ROUGE1 of the proposed CAG-RAG model. These improvements on several metrics attest to the system's flexibility in different subdomains of clinical text. Figure 4 presents the scores of the evaluation metrics of the proposed CAG-RAG model with respect to the existing DALL-M architecture [10] in terms of diagnosis and medication metrics.

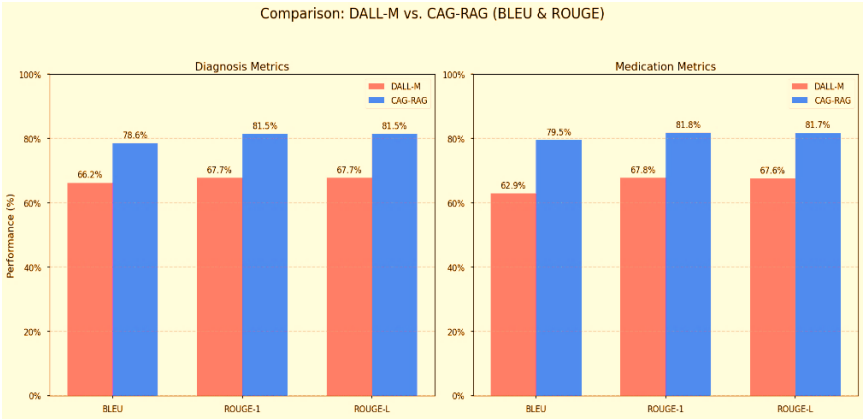

Figure 4. Diagnosis and Medication metrics DALL-M and CAG-RAG Architecture

Table 1. Improvements in CAG-RAG model (BLEU & ROUGE)

Diagnosis Metrics	**DALL - M**	**CAG-RAG**	**Performance Improvement**
BLEU	66.2	78.6	15.7 %
ROUGE -1	0.67	0.81	17 %
ROUGE -L	0.67	0.81	17 %
Medication Metrics	**DALL - M**	**CAG-RAG**	**Performance Improvement**
BLEU	62.9	79.5	20.8 %
ROUGE -1	0.67	0.82	17.1 %
ROUGE -L	0.67	0.81	17.3 %

5. Conclusion and Future Work

A GAN-based system is used to create synthetic clinical text data to overcome problems such as data scarcity, privacy, and class imbalance. The procedure involved are text preprocessing, topic modeling using Latent Dirichlet Allocation and iterative training of the generator and discriminator of the GAN. The generated data closely replicated real clinical text, which was useful for dataset augmentation in healthcare AI systems. This synthetic data was used in a Retrieval-Augmented Generation configuration to train a Large Language Model to generate context-aware responses to clinical questions. Incorporating the data in a Chatbot interface proved its applied relevance in clinical decision support and AI facilitated healthcare applications.

The work can be extended by incorporating further innovative NLP approaches, like transformer-based models, in order to enhance the cohesiveness and variability of produced text. Domain knowledge can be also added with the help of pre-trained medical embedding to medical datasets to improve synthetic data even further. Using the data generated for implementation in clinical decision support systems, training more sophisticated machine learning algorithms, and streamlining the metric evaluation with expert domain feedback, will make sure that the synthetic data is consistent with real healthcare standards, thereby offering a groundwork for clinical AI systems that can be both efficacious and reliable.

CRediT Author Contribution Statement

Nasreen Fathima: Conceptualization, Methodology, Data curation, Software Implementation, Analysis, Writing- original draft; Kavitha Ganesh: Conceptualization, Methodology, Writing- Review & editing, Visualization, Supervision, Project administration.

References

[1] Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang and Xia Hu, "Large Language Models for Healthcare Data Augmentation: An Example on Patient-Trial Matching", in *Proceedings of AMIA Annual Symposium 2023 (AMIA 2023)*, 11-15 November 2023, New Orleans, USA, Print ISSN: 1559-4076, Online ISSN: 1942-597X, pp. 1324-1333, Published by AMIA, DOI: 10.3390/healthcare12081147, Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10785941/pdf/1147.pdf.

[2] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar *et al.*, "A Comprehensive Overview of Large Language Models", *ACM Transactions on Intelligent Systems and Technology*, Online ISSN: 2157-6904, Vol. 16, No. 5, 18 August 2025, pp. 1–72. Published by Association for Computing Machinery, DOI: 10.1145/3744746, Available: https://dl.acm.org/doi/10.1145/3744746.

[3] Dandan Wang and Shiqing Zhang, "Large Language Models in Medical and Healthcare Fields: Applications, Advances, and Challenges", *Artificial Intelligence Review*, Print ISSN: 0269-2821, Online ISSN: 1573-7462, Vol. 57, 20 September 2024, pp. 1–48, Published by Springer, DOI: 10.1007/s10462-024-10921-0, Available: https://link.springer.com/article/10.1007/s10462-024-10921-0.

[4] Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib *et al.*, "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges", *IEEE Access*, Online ISSN: 2169-3536, Vol. 12, 13 February 2024, pp. 26839–26874, Published by IEEE, DOI: 10.1109/ACCESS.2024.3365742, Available: https://ieeexplore.ieee.org/document/10433480.

[5] Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L. Chaunzwa, Idalid Franco *et al.*, "Large Language Models to Identify Social Determinants of Health in Electronic Health Records", *NPJ Digital Medicine*, Online ISSN: 2398-6352, Vol. 7, 11 January 2024, pp. 1–14, Published by Nature, DOI: 10.1038/s41746-023-00970-0, Available: https://www.nature.com/articles/s41746-023-00970-0.

[6] Rumeng Li, Xun Wang and Hong Yu, "Two Directions for Clinical Data Generation with Large Language Models: Data-to-Label and Label-to-Data", in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6-10 December 2023, Singapore, Online ISSN: 2307-387X, pp.7129–7143, Published by Association for Computational Linguistics, DOI: 10.18653/v1/2023, Available: https://aclanthology.org/2023.findings-emnlp.474/.

[7] David M. Blei, Andrew Y. Ng and Michael I. Jordan, "Latent Dirichlet Allocation", *Annals of Journal of Machine Learning Research (JMLR)*, Print ISSN: 1532-4435, Vol. 3, January 2003, pp. 993–1022, Published by JMLR, DOI: 10.5555/944919.944937, Available: https://www.jmlr.org/papers/v3/blei03a.html.

[8] Hanna M. Wallach, David M. Mimno and Andrew McCallum, "Rethinking LDA: Why Priors Matter", *in Proceedings of the 23rd International Conference on Neural Information Processing Systems (NeurIPS 2009)*, 7-12 December

2009, Vancouver, Canada, ISBN: 978-1-61567-911-9, pp. 1973–1981, Published by Curran Associates Inc, DOI: 10.5555/2984093.2984314. Available: https://papers.nips.cc/paper/3854-rethinking-lda-why-priors-matter.html.

[9] Hawraa Ali Taher, Noralhuda N. Alabidand and Bushra Mahdi Hasan, "Integration Named Entity Recognition and Latent Dirichlet Allocation to Enhance Topic Modeling", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, Vol. 9, No. 2, pp. 20–30, 1 April 2025, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2025.02.002, Available: http://aetic.theiaer.org/archive/v9/v9n2/p2.html.

[10] Chihcheng Hsieh, Catarina Moreira, Isabel Blanco Nobre, Sandra Costa Sousa and Chun Ouyang *et al*., "DALL-M: Context-Aware Clinical Data Augmentation with LLMs", *Computers in Biology and Medicine*, ISSN: 0010-4825, Vol. 190, May 2025, pp. 1–15, Published by Elsevier, DOI: 10.1016/j.compbiomed.2025.110022, Available: https://www.sciencedirect.com/science/article/pii/S0010482525003737.

[11] Atif Latif and Jihie Kim, "Evaluation and Analysis of Large Language Models for Clinical Text Augmentation and Generation", *IEEE Access*, Online ISSN: 2169-3536, Vol. 12, 3 April 2024, pp. 48987–48996, Published by IEEE, DOI: 10.1109/ACCESS.2024.3384496, Available: https://ieeexplore.ieee.org/document/10489969.

[12] Dong-Jin Kim, Tae-Hyun Oh, Jinsoo Choi and In So Kweon, "Semi-Supervised Image Captioning by Adversarially Propagating Labeled Data", *IEEE Access*, Online ISSN: 2169-3536, Vol. 12, 5 July 2024, pp. 93580–93592, Published by IEEE , DOI: 10.1109/ACCESS.2024.3423790, Available: https://ieeexplore.ieee.org/document/10586974.

[13] Hamideh Ghanadian, Isar Nejadgholi and Hussein Al Osman, "Socially Aware Synthetic Data Generation for Suicidal Ideation Detection using Large Language Models", *IEEE Access*, Online ISSN: 2169-3536, Vol. 12, 24 January 2024, pp. 14350–14363, Published by IEEE, DOI: 10.1109/ACCESS.2024.3358206, Available: https://ieeexplore.ieee.org/document/10413447.

[14] Saman Motamed, Patrik Rogalla and Farzad Khalvati, "Data Augmentation using Generative Adversarial Networks for GAN-Based Detection of Pneumonia and COVID-19 in Chest X-Ray Images", *Annals of Informatics in Medicine Unlocked*, Print ISSN: 2352-9148, Vol. 27, 22 November 2021, pp. 1–7, Published by Elsevier, DOI: 10.1016/j.imu.2021.100779, Available: https://www.sciencedirect.com/science/article/pii/S2352914821002501.

[15] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng and Xiahui Jiang *et al*., "Latent Dirichlet Allocation and Topic Modeling: Models, Applications: A Survey", *Multimedia Tools and Applications*, Online ISSN: 1380-7501,Vol. 78, 28 November 2018, pp. 15169–15211, Published by Springer Nature, DOI: 10.1007/s11042-018-6894-4, Available: https://link.springer.com/article/10.1007/s11042-018-6894-4.

[16] Anmol Arora and Ananya Arora, "Generative Adversarial Networks and Synthetic Patient Data: Current Challenges and Future Perspectives", *Future Healthcare Journal*, Print ISSN: 2514-6645, Online ISSN: 2514-6653, Vol. 9, No. 2, 31 July 2022, pp. 190–193, Published by Elsevier, DOI: 10.7861/fhj.2022-0013, Available: https://www.sciencedirect.com/science/article/pii/S2514664524005009.

[17] Angona Biswas, Md Abdullah Al Nasim, Al Imran, Anika Tabassum Sejuty, Fabliha Fairooz *et al*., "Generative adversarial networks for data augmentation", in *Data Driven Approaches on Medical Imaging*, Singapore: Springer Nature, 17 October 2023, Online ISBN: 978-3-031-47772-0, Print ISBN: 978-3-031-47771-3, ch. 8, pp. 159–177, DOI: 10.1007/978-3-031-47772-0_8, Available: https://link.springer.com/chapter/10.1007/978-3-031-47772-0_8.

[18] Sana Imtiaz, Muhammad Arsalan, Vladimir Vlassov and Ramin Sadre, "Synthetic and Private Smart Health Care Data Generation using GANs", in *Proceedings of the 30th International Conference on Computer Communications and Networks (ICCCN 2021)*, 19–22 July 2021, Athens, Greece, Online ISBN: 978-1-6654-1278-0, E-ISBN: 978-1-6654-4835-2, pp. 1–7, Published by IEEE, DOI: 10.1109/ICCCN52240.2021.9522203, Available: https://ieeexplore.ieee.org/document/9522203.

[19] Gleb Kumichev, Pavel Blinov, Yulia Kuzkina, Vasily Goncharov and Galina Zubkova *et al*., "Medsyn: LLM-based Synthetic Medical Text Generation Framework", in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2024)*, 15-19 September 2024, Vilnius, Lithuania, Online ISBN: 978-3-031-70381-2, E-ISBN: 978-3-031-70380-5, pp.215–230, Published by Springer, DOI: 10.1007/978-3-031-70381-2_14, Available: https://link.springer.com/chapter/10.1007/978-3-031-70381-2_14.

[20] Pablo Picazo-Sanchez and Lara Ortiz-Martin, "Analysing the Impact of ChatGPT in Research", *Applied Intelligence*, Print ISSN: 0924-669X, Online ISSN: 1573-7497, Vol. 54, 21 March 2024, pp. 4172–4188, Published by Springer, DOI: 10.1007/s10489-024-05298-0, Available: https://doi.org/10.1007/s10489-024-05298-0.