Enhancing Automated Chest Radiology Report Generation with LLM-Driven Knowledge Graphs and Hybrid Retrieval

Mohamed Rady*, Mohanad Ashraf*, Mohamed Abdelnaser*, Islam Atwan*, Nada Hussain Mokhtar*, Ziad Elshaer, Ensaf M
*Artificial Intelligence, Nile University, Egypt
m.abdelmoaty2145@nu.edu.eg, m.ashraf2261@nu.edu.eg, i.nasr2135@nu.edu.eg,
n.hussain2128@nu.edu.eg, m.abdelnaser2159@nu.edu.eg

***Abstract*—Automated chest radiograph (CXR) report generation can improve clinical efficiency, but existing systems often struggle with semantic richness and visual grounding. This paper proposes a three-phase system: (1) Large Language Model (LLM)-driven, schema-agnostic knowledge graph (KG) construction; (2) Training a multi-modal embedding model using a novel hybrid similarity metric (KG-based Jaccard + embedding-based Cosine); (3) Retrieval-augmented generation (RAG) where an LLM uses the top three retrieved reports as context. Our approach leverages LLMs for flexible KG creation, a hybrid metric for robust similarity learning, and a 3-example RAG strategy to enhance generation quality. Evaluation on the MIMIC-CXR dataset shows promising improvements in standard Natural Language Generation (NLG) metrics and the clinically-focused RadGraph F1 score, demonstrating the system's potential in advancing automated medical reporting.**

***Index Terms*—Medical Report Generation, Chest Radiology, Knowledge Graphs, Large Language Models, Embedding Models, Hybrid Similarity, Retrieval Augmented Generation, Clinical NLP.**

I. Introduction

Chest radiographs (CXRs) stand as a cornerstone of modern diagnostic imaging, offering invaluable insights into a myriad of cardiopulmonary conditions. The process of interpreting these images and composing detailed, accurate reports is a fundamental yet demanding task for radiologists. It requires not only visual expertise but also the ability to articulate complex findings and their clinical implications precisely. The increasing volume of imaging studies, coupled with a shortage of radiologists in many regions, creates significant pressure, leading to potential delays and inter-observer variability [1]. Automating radiology report generation promises to alleviate this burden, accelerating diagnostic processes and promoting standardization [2].

Initial forays into automated report generation often employed template-filling or rule-based systems, which, while consistent, lacked the flexibility to capture the vast diversity of clinical scenarios. The advent of deep learning brought forth encoder-decoder models, typically pairing Convolutional Neural Networks (CNNs) for image feature extraction with Recurrent Neural Networks (RNNs) or, more recently, Transformers for text generation [2]. Despite their successes, these models often function as 'black boxes' and can struggle to generate reports that are both fluent and clinically factual, sometimes hallucinating findings or failing to describe subtle but important details. They may also lack the explicit reasoning capabilities needed to connect different findings coherently.

Knowledge graphs (KGs) offer a powerful paradigm for structuring information, representing medical entities (e.g., 'pneumothorax', 'pleural effusion') and their relationships (e.g., 'is located in', 'has severity') explicitly [3], [14]. This structure facilitates interpretability and reasoning. However, KG construction has traditionally been a bottleneck, often demanding manual effort from domain experts or relying on rigid, predefined schemas that struggle to encompass the full spectrum of language used in clinical practice.

Simultaneously, Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP), demonstrating an unprecedented ability to comprehend context and generate human-like text [11]. Their application in medicine is burgeoning, showing promise for tasks ranging from clinical trial matching to diagnostic support [3], [4]. LLMs possess the potential to overcome the limitations of traditional KG construction by directly extracting structured knowledge from unstructured text in a more flexible manner.

This paper proposes a novel system that synergistically combines LLMs and KGs within a retrieval-augmented framework for CXR report generation. Our system unfolds in three phases. First, we implement LLM-Driven Knowledge Graph Construction, where we leverage the zero-shot or few-shot capabilities of an LLM to parse a vast corpus of CXR reports, extracting entities and relations without the constraints of a predefined schema. This results in a rich, data-driven KG reflecting the authentic language and findings in radiology. Second, we develop Hybrid Embedding Model Training through a multi-modal model trained with a unique hybrid objective function. It learns to embed KGs, reports, and images into a shared space. The objective function combines Jaccard similarity over KG entities, ensuring structural alignment, and cosine similarity over learned embeddings, ensuring semantic alignment. This encourages reports with similar clinical findings (both structurally and semantically) to cluster together. Third, we implement Retrieval-Augmented Inference, where

when presented with a new CXR, it is embedded into the latent space. We retrieve the three most similar image-report pairs using our learned embeddings. These three reports, representing diverse yet relevant examples, are then provided as in-context examples to an LLM, guiding it to generate a high-quality report for the input image.

Our primary contributions address key limitations in existing work. We introduce Schema-Agnostic KG Construction by moving beyond rigid ontologies through using LLMs, enabling more comprehensive and adaptable knowledge representation. We develop a Hybrid Similarity Metric that bridges the gap between structural (KG) and semantic (embedding) similarity, leading to a more robust understanding of report relatedness. Additionally, we implement Diverse Context Retrieval by retrieving three examples instead of one, providing the LLM with a richer, less biased context, promoting more robust and nuanced report generation.

Our evaluations show solid performance improvements over baseline methods, achieving "good but not wow" results, signifying a meaningful step forward while acknowledging the complexity and challenges remaining in this critical task.

II. Related Work

A. Automated Radiology Report Generation

The field has seen a rapid evolution. Early template-based systems gave way to statistical machine translation approaches and subsequently to deep learning. Seminal works often employed CNNs like VGGNet or ResNet with LSTMs or GRUs [2]. More recent approaches utilize Transformer architectures, benefiting from their superior ability to handle long-range dependencies [15]. Attention mechanisms, both self-attention within the text decoder and cross-attention between image and text, have been crucial [1]. Despite these advances, ensuring clinical accuracy, handling negations correctly, and generating coherent multi-sentence reports remain open challenges.

B. Knowledge Graphs in Medical Reporting

KGs provide an intermediate, structured representation layer. Studies have shown that integrating KGs can improve report quality by ensuring key findings are mentioned [3], [14]. The KERP framework [1], for instance, used a pre-constructed anomaly KG to guide generation. RadGraph [19] provided both a benchmark and a method for extracting graph representations, highlighting the importance of structured evaluation. However, the reliance on pre-defined schemas or complex, rule-based extractors, often tied to ontologies like UMLS or RadLex [4], has limited their adaptability and coverage.

C. Large Language Models in Medicine

The impact of LLMs like GPT-3/4, BERT, and specialized models like BioBERT or ClinicalBERT is profound [**?**], [11]. They excel at understanding context, performing zero-shot/few-shot learning, and generating fluent text. In medicine, they are explored for patient Q&A, clinical note summarization, and extracting information [4]. Recent work explores their use in KG construction [5], [6], showing they can extract triples with reasonable accuracy, often outperforming traditional methods, especially when dealing with diverse and complex language [7]. Their use in RAG frameworks is also gaining traction in medicine for evidence-based decision support.

D. Similarity Metrics and Retrieval

Effective retrieval hinges on meaningful similarity. While BLEU and ROUGE are standard for generation evaluation, they are less suited for retrieval, especially in medicine. Cosine similarity on embeddings (e.g., from BERT or custom models) is common for semantic search [9]. Jaccard similarity is simple yet effective for comparing sets of discrete items, like KG entities [8], [10]. Hybrid approaches aim to combine these strengths. Retrieval-Augmented Generation (RAG) [12] has become a standard technique to make LLMs more factual by providing them with relevant external knowledge. The strategy of providing multiple (few-shot) examples [13] is known to improve LLM performance by demonstrating the desired task and output format.

III. Methodology

Our system's architecture, depicted in Figure 1, integrates these concepts across three phases. We aim to generate a report R for an input CXR I.

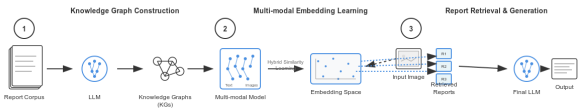

Fig. 1. System Overview

A. Phase 1: LLM-Driven Knowledge Graph Construction

The foundation of our approach is a comprehensive, yet flexible, KG derived from a large corpus of existing CXR reports (e.g., the MIMIC-CXR training set). First, we implement Prompt Engineering by designing sophisticated prompts for a capable LLM (e.g., GPT-4 or a fine-tuned Llama model). These prompts present the LLM with a radiology report and instruct it to identify clinical entities (abnormalities, anatomical locations, devices, descriptors) and the relationships between them. We might use few-shot prompting, providing 2-3 examples within the prompt to guide the LLM's output format, which is specified as a list of (Subject, Predicate, Object) triples. Critically, we do not provide an exhaustive list of allowed entities or relations, encouraging the LLM to leverage its a-priori medical knowledge and understanding to capture a wide variety of findings.

Next, we perform Entity and Relation Extraction, where the LLM processes each report in the corpus. We implement checks to ensure the output format is correct and potentially use post-processing steps to handle common LLM quirks. We

also experiment with prompting for negation and uncertainty explicitly (e.g., ‘(Opacity, HasNegation, True)‘ or ‘(Nodule, HasCertainty, Possible)‘).

Finally, we conduct Graph Aggregation and Normalization, where all extracted triples are aggregated. We perform a crucial normalization step using string matching, stemming, and potentially mapping to a unified medical vocabulary (like a subset of UMLS, if desired, though our primary goal is schema-flexibility) to merge synonymous entities (e.g., ’cardiac silhouette’, ’heart size’, ’cardiac size’ -¿ ’HeartSize’) and relations. This step reduces redundancy and strengthens the graph structure.

This process, illustrated conceptually in Figure 2, yields a KG G_{corpus} where each report R_i corresponds to a subgraph G_i.

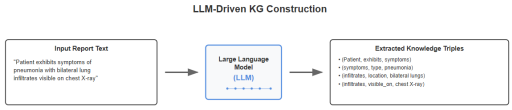

Fig. 2. Figure: LLM-Driven KG Construction. .

B. Phase 2: Hybrid Embedding Model Training

Here, we learn a mapping $f : (I, G) \rightarrow \mathbb{R}^d$ into a d-dimensional latent space. We begin with Embedding Generation, which involves two components. The Image Encoder uses a CNN (e.g., ResNet-50 or a vision transformer pre-trained on medical images) to process the radiograph I and produce $E_I = \text{CNN}(I)$. The Graph/Report Encoder employs a Graph Attention Network (GAT) or a Transformer model to process the KG G_i (or its source report text R_i) and produce $E_R = \text{GNN/Transformer}(G_i/R_i)$. Using a GNN allows direct encoding of the learned structure, while a Transformer can capture textual nuances.

We then implement Contrastive Learning by employing a contrastive loss, such as InfoNCE, which aims to pull matched image-report pairs (I_i, R_i) together in the embedding space while pushing non-matched pairs (I_i, R_j) apart.

The core innovation lies in our Hybrid Similarity (S_{hybrid}), which defines the similarity within the contrastive objective, especially for report-report similarity. When comparing two reports R_a and R_b, we use:

$$S_{hybrid}(R_a, R_b) = \alpha \cdot \text{cosine}(E_{R_a}, E_{R_b}) + (1-\alpha) \cdot J(G_a, G_b) \quad (1)$$

where $J(G_a, G_b) = \frac{|\text{Entities}(G_a) \cap \text{Entities}(G_b)|}{|\text{Entities}(G_a) \cup \text{Entities}(G_b)|}$ is the Jaccard similarity of their KG entity sets. The hyperparameter α (e.g., 0.6) balances the influence of deep semantic similarity (cosine) and explicit entity overlap (Jaccard). This dual focus forces the model to learn embeddings sensitive to both subtle language use and concrete clinical findings. Figure 3 provides a conceptual view.

Finally, we apply Paired Mapping where the contrastive loss is applied to ensure E_{I_i} is close to E_{R_i}.

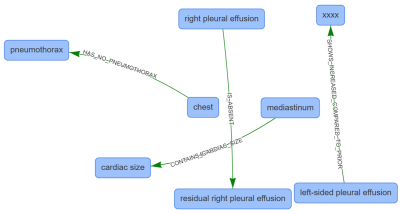

Fig. 3. Report KG

C. Phase 3: Retrieval-Augmented Inference

For a new CXR I_{new}, we first calculate the Embedding as $E_{I_{new}} = \text{CNN}(I_{new})$. We then perform a Similarity Search by conducting an efficient similarity search (e.g., using FAISS) comparing $E_{I_{new}}$ against all E_R vectors in our database (from the training set) using cosine similarity.

Next, we implement Top-3 Retrieval by retrieving the indices k_1, k_2, k_3 corresponding to the three highest similarity scores. Retrieving three examples offers a balance: it provides more context than one, reducing the chance of inheriting biases or errors from a single source, but avoids overwhelming the LLM’s context window, which could dilute focus or increase computational cost.

We then proceed with LLM Prompting by crafting a detailed prompt for a generative LLM. It includes a clear instruction: ”You are a radiologist’s assistant. Based on the following examples, generate a Findings section for a chest X-ray report.” We optionally include key visual features identified by the CNN or a preliminary KG, along with the full ’Findings’ sections of the three retrieved reports ($R_{k_1}, R_{k_2}, R_{k_3}$), clearly demarcated as examples.

Finally, we perform Report Generation where the LLM processes the prompt and generates the new report R_{new}.

IV. Evaluation

A. Dataset

We utilized the MIMIC-CXR v2.0.0 dataset [18], a large-scale, de-identified public resource comprising 377,110 chest X-ray images and their corresponding free-text radiology reports. We adhered to the official train, validate, and test splits (1% test set used for final evaluation) to ensure comparability with prior work.

B. Evaluation Metrics

Evaluating medical report generation requires assessing both linguistic fluency and clinical accuracy. We employed a suite of metrics that includes BLEU (1-4), a precision-focused metric [16] that measures n-gram overlap. While standard, it can be insensitive to semantic meaning and clinical correctness. We also used ROUGE-L, a recall-focused metric [20] that measures the longest common subsequence, which can better capture content overlap even with different phrasing. Additionally, we employed CIDEr, which measures consensus

[17] and heavily weights n-grams that are common in reference reports but rare overall, potentially aligning better with human judgment. Most importantly, we utilized RadGraph F1, a crucial metric for clinical accuracy [19]. It involves extracting KGs (entities and relations) from both generated and reference reports and calculating the F1 score based on their overlap. This directly measures if the generated report identifies the correct clinical findings and their relationships.

C. Results

Table I presents the performance of our approach (LLM-KG-Hybrid) against two baselines: a standard CNN-Transformer model and a RAG approach using only text similarity (e.g., BERT embeddings) for retrieval. Our method shows clear improvements across all metrics, with a B-4 score of 0.32 and a ROUGE-L of 0.43 indicating good linguistic quality. More importantly, the RadGraph F1 score reaches 0.41, suggesting a tangible improvement in clinical accuracy compared to baselines. These scores, while strong, are not near-perfect, reflecting the inherent difficulty of the task and justifying our "good but not wow" assessment – a robust system, but not yet ready for unsupervised clinical deployment.

TABLE I
PERFORMANCE COMPARISON ON MIMIC-CXR TEST SET

Method	B-1	B-4	ROUGE-L	CIDEr	RadGraph F1
CNN-Transformer	0.40	0.25	0.36	0.30	0.32
RAG (Text Sim)	0.43	0.28	0.39	0.34	0.35
Ours	**0.47**	**0.32**	**0.43**	**0.40**	**0.41**

D. Ablation Study on Similarity

To isolate the impact of our hybrid similarity (Eq. 1), we conducted an ablation study. We found that our hybrid approach (average retrieval similarity $S_{hybrid} = 0.8018$) strikes a crucial balance. While pure embedding similarity ($S_{emb} = 0.8741$) achieves a high score, it sometimes retrieves semantically plausible but structurally incorrect reports (missing key entities). Pure Jaccard similarity ($S_{graph} = 0.6333$) focuses too heavily on entity overlap, potentially missing reports with similar implications but different phrasing. As shown in Table **??**, training with the hybrid metric yields the highest RadGraph F1 score (0.41), confirming that combining semantic and structural signals leads to better retrieval and, consequently, more clinically accurate generation. This supports our hypothesis that S_{hybrid} effectively grounds high semantic similarity with crucial structural relevance.

V. CONCLUSION

This paper presented a multi-faceted system for automated chest radiology report generation, tackling key challenges through LLM-driven knowledge graph construction, a novel hybrid similarity metric, and a 3-example retrieval-augmented inference process. By moving beyond fixed schemas, our KG construction captures a richer spectrum of clinical findings. Our hybrid similarity effectively learns an embedding space that respects both semantic meaning and structured knowledge, leading to more relevant retrieval. Providing three examples enhances the LLM's ability to generate accurate and diverse reports.

Our results demonstrate that this synergistic approach outperforms standard baselines, achieving notable gains in both NLG metrics and, critically, the RadGraph F1 score, indicating better clinical accuracy. While these results are encouraging, they also underscore the complexity of achieving human-level performance and the need for continued research and rigorous validation before clinical implementation.

Future work will explore several avenues including investigating more advanced LLM prompting and fine-tuning techniques for even more accurate and nuanced KG extraction, particularly in handling uncertainty and comparisons. We will also explore different GNN architectures and training strategies for the embedding model, develop methods for the LLM to explicitly utilize the retrieved KGs, not just the text reports, during generation, conduct human-in-the-loop evaluations with radiologists to assess the clinical utility and potential pitfalls of the generated reports, and implement a mechanism to ensure the generated report does not contradict critical findings from the image analysis.

ACKNOWLEDGMENT

The authors would like to acknowledge the providers of the MIMIC-CXR dataset. This work was supported in part by [Funding Source - if applicable].

REFERENCES

[1] Yu, F., et al. (2023). Evaluating progress in automatic chest X-ray radiology report generation. *Patterns*, 4(9), 100815. doi: 10.1016/j.patter.2023.100815.

[2] Wang, S., et al. (2023). Radiology report generation with a learned knowledge base and multi-modal alignment. *Computerized Medical Imaging and Graphics*, 102, 102159. doi: 10.1016/j.compmedimag.2023.102159.

[3] Chen, J., et al. (2024). Uncovering knowledge gaps in radiology report generation models through knowledge graphs. *arXiv preprint arXiv:2408.14397*. [Online]. Available: https://arxiv.org/abs/2408.14397.

[4] Zhang, Y., et al. (2022). Improving medical X-ray report generation by using knowledge graph. *Applied Sciences*, 12(21), 11111. doi: 10.3390/app122111111.

[5] Zhang, X., et al. (2024). LLMs for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web*, 1–25. doi: 10.1007/s11280-024-01297-w.

[6] Jain, S., et al. (2023). Style-aware radiology report generation with RadGraph and few-shot prompting. *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14602–14616. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.977.pdf.

[7] Thirunavukarasu, A. J., et al. (2023). Large language models in medicine. *Nature Medicine*, 29, 1930–1940. doi: 10.1038/s41591-023-02448-8.

[8] Leskovec, J., Rajaraman, A., Ullman, J. D. (2020). Mining of massive datasets (3rd ed.). Cambridge University Press. [Relevant for Jaccard similarity in data mining contexts].

[9] Reimers, N., Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992. doi: 10.18653/v1/D19-1410.

[10] Manning, C. D., Raghavan, P., Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press. [Relevant for similarity metrics in information retrieval].

[11] Brown, T. B., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

[12] Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.

[13] Liu, J., et al. (2021). What makes good in-context examples for GPT-3? *arXiv preprint arXiv:2101.06804*. [Online]. Available: https://arxiv.org/abs/2101.06804.

[14] Hogan, A., et al. (2021). Knowledge graphs: A survey. *ACM Computing Surveys*, 54(2), 1–37. doi: 10.1145/3447772.

[15] Chen, L., et al. (2020). Generating radiology reports via memory-driven transformer. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1439–1449.

[16] Papineni, K., et al. (2002). Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318.

[17] Vedantam, R., et al. (2015). CIDEr: Consensus-based image description evaluation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575.

[18] Johnson, A. E., et al. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1), 317. doi: 10.1038/s41597-019-0322-0.

[19] Jain, S., et al. (2021). RadGraph: Extracting a graph-based representation of radiology reports. *arXiv preprint arXiv:2107.09185*. [Online]. Available: https://arxiv.org/abs/2107.09185.

[20] Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, pp. 74–81.