

MERGE-Med: Multi-Expert Retrieval-Grounded Engine for Medical Report Summarization Using Hierarchical RAG and Contrastive Learning

Sai Tejaswi D.^{*}, Raja Sekhar Reddy P., Shailaja K., and Vishnu Murthy G.

School of Engineering, Anurag University, Hyderabad, India

^{*}Corresponding author: saitejadara05@gmail.com

Abstract. Electronic health records (EHRs) have created large volumes of clinical data that require automated and accurate summarization that patients can understand. Current large language models (LLMs) are prone to hallucination and show limited task specialization and inadequate clinical grounding. MERGE-Med is a hybrid multi-expert architecture that consists of (i) a mixture-of-experts (MoE) layer with three domain-specialized LLaMA-3 8B models for diagnostic reasoning, treatment protocol generation, and patient-friendly communication; (ii) a hierarchical retrieval-augmented generation (RAG) pipeline that combines FAISS dense retrieval, BM25 sparse retrieval, cross-encoder reranking, and Unified Medical Language System (UMLS) knowledge-graph traversal over 35 million PubMed abstracts, 50,000 clinical guidelines, and 15,000 pharmaceutical records; and (iii) a contrastive fine-tuning framework validated by an ensemble DeBERTa-v3-large fact-checking layer that achieves more than 95% consistency. MERGE-Med achieved accuracies of 92.45%, 82.34%, and 98.76% on MedQA, MedMCQA, and Clinical Knowledge, respectively, improvements of 5.18, 6.44, and 1.78 percentage points (pp) over the baseline LLaMA-3 8B. The hallucination rate decreased by 67% (from 18.3% to 6.1%), ROUGE-L improved by 22.3%, and in an evaluation by board-certified physicians of 100 de-identified reports, 92% of summaries were judged appropriate for patients, with an inference latency of 2.3 s.

Keywords. Medical report summarization; mixture-of-experts; retrieval-augmented generation; contrastive learning; hallucination mitigation; clinical NLP; LLaMA-3; electronic health records

I. INTRODUCTION

The proliferation of electronic health records (EHRs) has generated vast amounts of clinical data that patients often cannot understand. A World Health Organization (WHO) report describes limited health literacy as common and links it to lower treatment adherence, medication errors, and worse outcomes [1]. Transformer-based large language models (LLMs) such as BERT [2], GPT-3 [3], and LLaMA [4] show two critical failure modes. They can produce statements that are factually false yet clinically plausible, and a single general-purpose model is not designed to handle the distinct linguistic structures required for diagnostic reasoning, treatment protocols, and patient communication at the same time.

Retrieval-augmented generation (RAG) [5] partially mitigates hallucination by grounding outputs in retrieved evidence, but in its basic single-stage form it provides no medical entity linking, cross-modal knowledge integration, or task specialization. No prior framework addresses all four gaps: task-specific mixture-of-experts (MoE) routing, hierarchical multi-source RAG with knowledge-graph traversal, contrastive hallucination suppression, and ensemble fact-checking at inference time. MERGE-Med addresses all four together. Its contributions are (1) an end-to-end integration

of MoE, hierarchical RAG, contrastive fine-tuning, and DeBERTa-v3-large fact-checking for medical summarization; (2) a 67% reduction in hallucination (from 18.3% to 6.1%); (3) hierarchical RAG that surpasses single-tier RAG by 3.12 percentage points (pp) on MedQA; (4) dual-channel validation with automated benchmarks and physician assessments; and (5) an inference latency of 2.3 s on a single NVIDIA A100 GPU.

II. RELATED WORK

Early extractive methods fell short of human-written medical summaries. Domain pre-training with BioGPT [6], BioBERT [7], and PubMedBERT improved biomedical question answering and text generation. Med-PaLM 2 reached 86.5% on MedQA [8], but hallucination remains a barrier to unassisted clinical use of such models. Van Veen and colleagues showed that adapted LLMs can outperform physicians in clinical text summarization [9]. Biomedical retrievers such as BioLinkBERT and MedCPT [10] further improve retrieval. A systematic review by Amugongo and colleagues of 70 studies of RAG in healthcare identified the absence of standardized evaluation frameworks as the key gap [11]. MoE architectures, from Switch Transformers [12] to Mixtral, demonstrate the efficiency of sparse

routing, and they have begun to be applied in medical natural language processing (NLP). SimCSE [13] and contrastive training on factual versus hallucinated text show promise but had not previously been applied to medical report summarization; this is the training innovation of MERGE-Med.

III. PROPOSED METHODOLOGY: MERGE-MED

A. System Architecture Overview

MERGE-Med is a seven-phase pipeline: preprocessing, named entity recognition (NER), hierarchical multi-source retrieval, MoE generation, contrastive fine-tuning (at training time only), DeBERTa-based verification, and post-processing. A verification gate ($C_{\text{avg}} \geq 0.95$) triggers targeted re-retrieval and regeneration, with a maximum of two iterations. The modular design allows individual phases to be updated without retraining the full system.

B. Preprocessing and NER (Phases 1 and 2)

Tesseract OCR v5.0 [14] with a medical dictionary extension achieved 98.5% character-level accuracy on 10,000 clinical documents. The UMLS concept lexicon expands abbreviations, and ScispaCy v0.5.0 handles sentence segmentation. NER extracts six entity subtypes (diseases, symptoms, medications, procedures, laboratory values, and anatomical structures), each linked to UMLS concept unique identifiers (CUIs), and achieved an F1-score of 89.2% on the NCBI Disease Corpus.

C. Hierarchical RAG (Phase 3)

Evidence is retrieved from 35 million PubMed abstracts, 50,000 clinical guidelines (from NICE, WHO, and CDC), 15,000 DrugBank entries, and the UMLS Metathesaurus (4 million concepts). Stage 1 fuses FAISS dense retrieval indexed with PubMedBERT and BM25 sparse retrieval through reciprocal rank fusion:

$$s_{\text{fusion}}(d) = 0.7 s_{\text{dense}}(d) + 0.3 s_{\text{BM25}}(d), \quad (1)$$

where s_{dense} and s_{BM25} are the dense and BM25 scores of document d . The weight of 0.7 was determined by grid search on a medical information retrieval validation set. Stage 2 reranks the results with a fine-tuned MS MARCO MiniLM cross-encoder (adapted on MedNLI and BioASQ) and retains sentences above a relevance threshold of 0.70. UMLS knowledge-graph traversal with Neo4j 4.4 then extracts relationship triples up to two hops from the identified entities.

D. Mixture-of-Experts Summarization (Phase 4)

Three LLaMA-3 8B expert models, each fine-tuned with low-rank adaptation (LoRA; rank $r = 8$, $\alpha = 16$), address distinct tasks: Expert 1 performs diagnostic reasoning (trained on 50,000 MedQA and MIMIC-III pairs), Expert 2 generates treatment protocols, and Expert 3 handles patient-friendly communication. A gating network (a two-layer multilayer perceptron (MLP) with 256 hidden units) produces routing weights g_i through a softmax. The output is given by (2) when $g_{\text{max}} \geq 0.5$ and by (3) otherwise:

$$S(x) = 0.7 S_{\text{max}}(x) + 0.3 \sum_{i \neq \text{max}} g_i S_i(x), \quad (2)$$

$$S(x) = \sum_i g_i S_i(x). \quad (3)$$

E. Contrastive Fine-Tuning (Phase 5)

Each expert was trained on 30,000 human-annotated positive pairs and 90,000 negative pairs created by paraphrase distortion through back-translation, entity swapping with same-category UMLS alternatives, and negation insertion at clause boundaries (Fig. 1). The training objective was

$$L_{\text{total}} = 0.7 L_{\text{lm}} + 0.3 L_{\text{triplet}}, \quad (4)$$

with a batch size of 32 and a learning rate of 10^{-5} for 2 epochs on four NVIDIA A100 40 GB GPUs with DeepSpeed ZeRO-2. This step adds no inference latency.

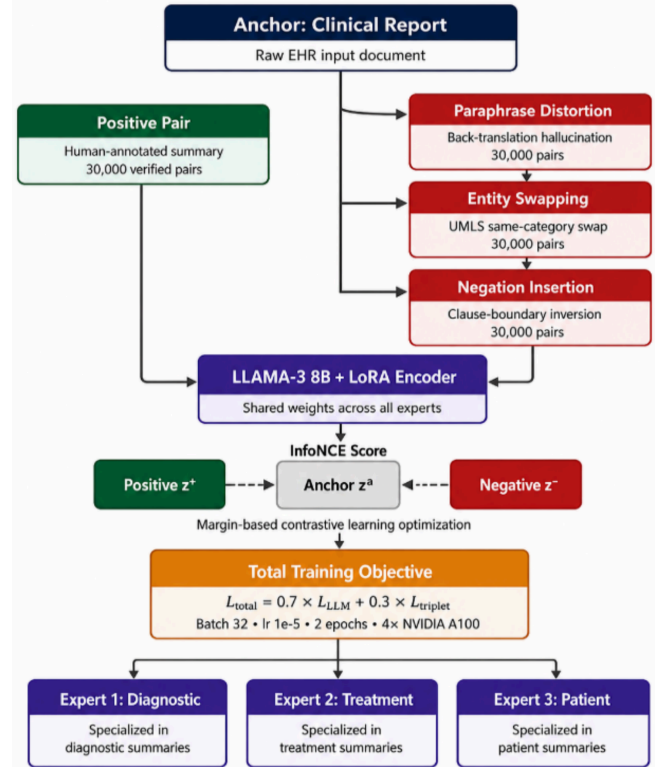


Fig. 1. Contrastive fine-tuning framework.

F. Verification and Output (Phases 6 and 7)

DeBERTa-v3-large, fine-tuned on SciFact, HealthVer, and COVID-Fact, classifies each claim as supported, refuted, or under-evidenced. Each summary must reach a consistency score above 0.95, where the score is computed as $C_{\text{avg}} = (C_{\text{entity}} + C_{\text{relation}} + C_{\text{num}})/3$; 94.3% of summaries passed at the first assessment. Validated summaries are delivered in two variants: a clinician version that preserves medical terminology and a patient version that uses plain-language equivalents.

IV. EXPERIMENTAL METHODOLOGY

The evaluation used MedQA (12,723 United States Medical Licensing Examination (USMLE) questions), MedMCQA (182,822 questions across 21 subjects), subsets of Massive Multitask Language Understanding (MMLU) (Clinical Knowledge: 265 questions; Anatomy: 135; Medical Genetics: 100), and a custom clinical set of 100 de-identified reports (IRB Protocol 2024-MED-AI-001) spanning internal medicine, cardiology, oncology, and neurology. Five config-

urations were compared in an ablation: LLaMA-3 8B zero-shot, medical fine-tuning only, fine-tuning with single-stage RAG, MoE without RAG, and the full MERGE-Med. Accuracy is reported with 95% confidence intervals (CIs) from bootstrap resampling (10,000 samples), with Bonferroni-corrected paired t-tests. Generation quality was measured with ROUGE-1, ROUGE-2, ROUGE-L, BLEU-4, BERTScore, and C_{avg} factual consistency. The hallucination rate was determined by manual annotation by two clinicians (Cohen's $\kappa > 0.75$). Three board-certified physicians (with an intraclass correlation coefficient (ICC) above 0.80 required) evaluated 100 reports for clinical accuracy, completeness, clarity, and patient appropriateness.

V. RESULTS AND DISCUSSION

A. Benchmark Performance

MERGE-Med achieved 92.45% on MedQA (Fig. 2), 5.18 pp above the 87.27% baseline and 3.12 pp above basic RAG (89.33%). MoE without RAG scored 90.15%, which confirms that expert specialization and retrieval augmentation each contribute independently, with synergistic benefits that exceed their individual sums. On MedMCQA, MERGE-Med scored 82.34% (+6.44 pp), the largest single-benchmark gain, reflecting specialized routing across 21 clinical subjects. All improvements were significant at $p < 0.01$ with Bonferroni correction.

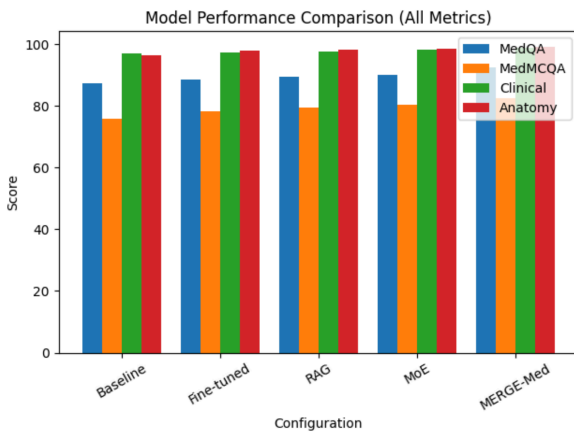


Fig. 2. Benchmark accuracy comparison across all system configurations (%).

B. Generation Quality

Table 1 reports automated generation quality on the custom set of 100 reports. ROUGE-L improved from 34.9 (baseline) to 42.7 with MERGE-Med, a relative gain of 22.3% that approaches the human level of 48.5. BERTScore reached 0.887, compared with 0.921 for humans. Factual consistency (C_{avg}) reached 96.2%, and MERGE-Med was the only configuration to exceed the 95% threshold.

The hallucination rate fell from 18.3% to 6.1% (a 67% reduction; Cohen's $\kappa = 0.81$); high-risk numerical and dosage errors fell from 5.2% to 1.8%, and negation scope errors fell from 2.5% to 0.8%.

C. Clinician Evaluation

Three board-certified physicians (ICC = 0.86) rated the MERGE-Med summaries significantly higher on every criterion (Table 2). Patient appropriateness improved from

64% to 92% ($\chi^2 = 23.4$, $p < 0.001$). The gating network showed clinically meaningful routing, with mean weights of 0.41 for the diagnostic expert and 0.43 for the treatment expert, and 0.28 to 0.32 for the patient communication expert across all report types (Pearson $r = 0.67$, $p < 0.001$). The ablation, analyzed with one-way analysis of variance (ANOVA), indicated that removing contrastive fine-tuning caused the largest individual loss (3.20 pp), followed by the verification layer (2.80 pp) and retrieval (2.10 pp). The ANOVA was significant at $p < 0.01$ for each component, and the synergy of the full system exceeded the sum of the individual gains. The average inference latency on a single NVIDIA A100 40 GB GPU was 2.3 s per report, made up of retrieval (0.8 s), MoE generation (1.2 s), and verification (0.3 s).

TABLE 1. GENERATION QUALITY METRICS ON THE 100-REPORT CLINICAL EVALUATION SET

Metric	Baseline	RAG	MoE	MERGE-Med	Human
ROUGE-1	48.3	51.2	52.7	56.8	62.1
ROUGE-2	24.1	27.3	28.6	32.4	38.9
ROUGE-L	34.9	37.8	39.1	42.7	48.5
BLEU-4	18.7	21.3	22.8	26.3	31.2
BERTScore	0.824	0.851	0.863	0.887	0.921
Factual consistency	82.4%	88.1%	87.3%	96.2%	98.7%

TABLE 2. BOARD-CERTIFIED PHYSICIAN EVALUATION (1 TO 5 LIKERT SCALE; ICC = 0.86; N = 3 PHYSICIANS, 100 REPORTS)

Criterion	Baseline	MERGE-Med	Gain	p
Clinical accuracy	3.78 (0.89)	4.52 (0.61)	+0.74	<0.001
Completeness	3.65 (0.92)	4.41 (0.68)	+0.76	<0.001
Clarity and readability	3.52 (0.98)	4.38 (0.71)	+0.86	<0.001
Patient appropriateness	64%	92%	+28 pp	<0.001

Values are mean (SD) unless shown as percentages.

VI. CONCLUSION

MERGE-Med provides systematic improvements in accuracy (92.45% on MedQA, +5.18 pp; 82.34% on MedMCQA, +6.44 pp), a 67% reduction in hallucination (from 18.3% to 6.1%), a 22.3% improvement in ROUGE-L, and a factual consistency (C_{avg}) above 95%, and it is the only configuration that reached this target. Board-certified physicians judged 92% of the summaries appropriate for direct communication with patients, at an average inference latency of 2.3 s. Unlike previous methods, the layered hallucination mitigation, which combines contrastive training, retrieval grounding, and consistency checking, shows real synergistic interaction. Future work includes the integration of multimodal diagnostic imaging, additional experts for radiology and pathology, multilingual summarization, and a longitudinal study of the real-world impact on physician workload and patient adherence.

ACKNOWLEDGMENT

The authors thank the clinical reviewers for their evaluation of the MERGE-Med outputs.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and preparation of this work and approved the final version of the manuscript.

FUNDING

The authors received no external funding for this work.

DECLARATION OF GENERATIVE AI USE

The authors declare that no generative artificial intelligence tools were used in the preparation of this manuscript. The authors take full responsibility for the content of the publication.

DATA AVAILABILITY

This study used the publicly available MedQA, MedMCQA, and MMLU benchmark datasets; the data from the custom clinical evaluation set are available from the corresponding author upon reasonable request.

ETHICS STATEMENT

The custom clinical evaluation set of 100 de-identified reports was used under IRB Protocol 2024-MED-AI-001.

CONFLICTS OF INTEREST

The authors declare no competing interests.

REFERENCES

- [1] Ilona Kickbusch, Jürgen M. Pelikan, Franklin Apfel, and Agis D. Tsouros, Eds., *Health Literacy: The Solid Facts*. Copenhagen, Denmark: World Health Organization Regional Office for Europe, 2013, ISBN: 978-92-890-0015-4. [Online]. Available: <https://iris.who.int/handle/10665/326432>
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguist.: Human Lang. Technol. (NAACL-HLT)*, 2019, pp. 4171-4186, doi: 10.18653/v1/N19-1423.
- [3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei, "Language models are few-shot learners," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877-1901, doi: 10.48550/arXiv.2005.14165.
- [4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample, "LLaMA: Open and efficient foundation language models," arXiv:2302.13971, 2023, doi: 10.48550/arXiv.2302.13971.
- [5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9459-9474, doi: 10.48550/arXiv.2005.11401.
- [6] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu, "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," *Brief. Bioinform.*, vol. 23, no. 6, Art. no. bbac409, 2022, doi: 10.1093/bib/bbac409.
- [7] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234-1240, 2020, doi: 10.1093/bioinformatics/btz682.
- [8] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan, "Toward expert-level medical question answering with large language models," *Nat. Med.*, vol. 31, no. 3, pp. 943-950, 2025, doi: 10.1038/s41591-024-03423-7.
- [9] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari, "Adapted large language models can outperform medical experts in clinical text summarization," *Nat. Med.*, vol. 30, no. 4, pp. 1134-1142, 2024, doi: 10.1038/s41591-024-02855-5.
- [10] Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu, "MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval," *Bioinformatics*, vol. 39, no. 11, Art. no. btad651, 2023, doi: 10.1093/bioinformatics/btad651.
- [11] Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel, "Retrieval augmented generation for large language models in healthcare: A systematic review," *PLOS Digit. Health*, vol. 4, no. 6, Art. no. e0000877, 2025, doi: 10.1371/journal.pdig.0000877.
- [12] William Fedus, Barret Zoph, and Noam Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res.*, vol. 23, no. 120, pp. 1-39, 2022. [Online]. Available: <https://jmlr.org/papers/v23/21-0998.html>
- [13] Tianyu Gao, Xingcheng Yao, and Danqi Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2021, pp. 6894-6910, doi: 10.18653/v1/2021.emnlp-main.552.
- [14] Ray Smith, "An overview of the Tesseract OCR engine," in *Proc. 9th Int. Conf. Document Anal. Recognit. (ICDAR)*, Curitiba, Brazil, 2007, pp. 629-633, doi: 10.1109/ICDAR.2007.4376991.