SAHA: Samvad AI for Healthcare Assistance

Aditya Kumar[1] **Rakesh Kumar Nayak**[1] **Janhavi Naik**[1]
Ritesh Kumar[1] **Dhiraj Bhatia**[2] **Shreya Agarwal**[1]
[1]Indian Institute of Information Technology, Surat, India
[2]Indian Institute of Technology, Gandhinagar, India
aditya.aimail@gmail.com
rknayak1947@gmail.com janhavi.141620@gmail.com
riteshkumar@iiitsurat.ac.in dhiraj.bhatia@iitgn.ac.in
shreya.agarwal@iiitsurat.ac.in

Abstract

This paper deals with the dual task of developing a medical question answering (QA) system and generating concise summaries of medical dialogue data across nine languages (English and eight Indian languages). The medical dialogue data focuses on two critical health issues: Head and Neck Cancer (HNC) and Cystic Fibrosis (NLP AI4health shared task). The proposed framework utilises a dual approach: a fine-tuned small Multilingual Text-to-Text Transfer Transformer (mT5) model for the conversational summarisation component and a fine-tuned Retrieval Augmented Generation (RAG) system integrating the dense intfloat/e5-large language model for the language-independent QA component. The efficacy of the proposed approaches is demonstrated by achieving promising precision in the QA task. Our framework achieved the highest F1 scores in QA for the three Indian languages, with F1 score of 0.3995 in Marathi, 0.7803 in Bangla, and 0.74759 in Hindi, respectively. We achieved the highest cometscore of 0.5626 on the Gujarati QA test set. For the dialogue summarisation task, our model registered the highest ROUGE-2 and ROUGE-L precision across all eight Indian languages, with English being the sole exception. These results confirm our approach potential to improve e-health in dialogue data for low-resource Indian languages.

1 Introduction

Understanding patient-centric medical dialogue systems is challenging due to multi-turn complexity, intent capture, cross-lingual semantics, and domain-specific terminologies (23; 10). The 2025 NLP-AI4Health shared task focuses on developing systems that can generate concise summaries and relevant responses to questions based on real-world medical conversations related to Head and Neck Cancer (HNC) and Cystic Fibrosis across 9 languages (8 Indian and English) (1). These data include multilingual and code-mixed interactions, emphasising the need for models that can generalise across languages. A significant dearth of high-quality annotated data for most Indian languages severely impedes the training and development of effective Indic medical dialogue understanding models. Addressing the need for understanding multilingual interactions and the high computational costs associated with developing such a system from scratch, we propose a framework to tackle such tasks effectively. For dialogue summarisation, we fine-tuned the Multilingual Text-to-Text Transfer Transformer (mT5) model (27) to generate coherent, domain-relevant summaries from patient-doctor dialogues. For question answering (QA), we designed a fined-tuned Retrieval-Augmented Generation (RAG) with given QA data pipeline that integrates a dense multilingual retriever based on the fine-tuned intfloat/e5-large model (16; 31), enabling language-independent and precise response generation. Our approach achieved promising results on both dialogue summarisation and QA tasks (see section 5). The paper is organised as follows: The following section highlights the existing work in QA and summarisation of medical dialogue data. Section 3 provides a detailed description of the proposed methodology. The results and discussion of the proposed approach are delineated in Section 5. Conclusion is present in Section 6 with some future points of action in Section 7.

2 Related Work

Medical dialogue summarisation and QA is a challenging task. In the Multi-turn conversational nature of patient–clinician interactions, extractive approaches often fail to capture relevant summaries (23; 10). Automatic evaluation is a major challenge in generative tasks such as dialogue response

Second Workshop on Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication, pages 80–85

and summarisation. Traditional generation tasks evaluation metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (2) frequently fall short in capturing semantic fluency and adequacy. To address this, newer metrics like BERTScore (19) and COMETscore (20), have been used for the evaluation of both tasks. Abstractive summarisation capable of paraphrasing content in its own words has gained prominence (17; 7; 8). The advent of pre-trained language models (PLMs) such as BioBERT (18) and ClinicalBERT (9) is a game-changer in the medical field (15). For generative tasks, sequence-to-sequence architectures, BART (17) and T5 (21) have become the standard choices. To mitigate hallucinations in PLMs, RAG (16) is introduced. The performance of the RAG pipeline depends heavily on the quality of its retriever. Our pipeline employs the multilingual model *intfloat/e5-large* (28). To address large-scale retrieval efficiently, we use Facebook AI Similarity Search (FAISS) (14).

3 Proposed Methodology

This section details the proposed architecture for the medical QA and dialogue summarisation task.

3.1 Dataset Overview

The text box below provides a snippet of a dialogue from the training dataset. Table 1 and Table 2 show the data statistcs.


```
{
  "speaker": "Health Worker",
  "date": "2025-10-05",
  "dialogue": "I'm Dr. Sen. I want to hear about your 2-month-old boy, his cough since
              birth, slow weight gain, and the oily stools you mentioned. Where are you
              traveling from today?"
},
{
  "speaker": "Patient",
  "date": "2025-10-05",
  "dialogue": "Navi Mumbai, actually. Hes been coughing since birth, mother says, and he
              seems to get tired after feeds. Weight isnt climbing fast, and stools look greasy
              ."
}
```


Each QA file has pairs as follows:


```
{
  "questions": [
    {
      "question": "Can you explain what the sweat test is and how reliable it is for cystic
                  fibrosis?",
      "answer": "The sweat test measures chloride levels in sweat and is the primary
                  diagnostic test for cystic fibrosis. High chloride levels support a CF
                  diagnosis, normal levels rule it out, and borderline results may require
                  repeated or genetic testing."
    }
  ]
}
```


3.2 Preprocessing

The initial pre-processing steps included language detection (4), text normalisation (6), sentence segmentation (5), stopword removal, and multilingual encoding (22), and duplicate dialogues deletion(11). We claim novelty in adding a query validation layer that checks the retrieved content. If similarity is below threshold, the system returns a safe fallback, preventing hallucination: *"The system does not have sufficient information to answer this question. Please consult a certified medical professional."*. All processed data and responses are cached for faster retrieval during evaluation. Metadata such as source, retrieval confidence, and language tags are stored with each instance. The final pre-processing steps included: Parsing and aligning dialogue–summary pairs using unique identifiers, removing null entries, normalising inconsistent speaker tags, spaces, and punctuation. The data is structured into the following fields: `id`, `dialogue`, and `summary`, stored as Hugging Face Dataset objects for efficient loading (29). The cleaned dataset is then cached to optimise runtime efficiency during fine-tuning.

Table 1: QnA task dataset statistics

Language	Train_QnA	Dev_QnA	Test_QnA
Assamese	204	1200	65
Bangla	5064	1200	78
Dogri	1548	1200	82
English	95696	1632	87
Gujarati	5004	1200	52
Hindi	13420	1232	86
Kannada	11496	1200	28
Marathi	8916	1200	85
Tamil	9092	1200	64
Telugu	26352	1200	68

Table 2: Dialogues for Summarisation data statistics

Language	train_dialogues	dev_dialogues	test_dialogues
Assamese	112,338	4372	4337
Bangla	433,832	6794	6365
Dogri	197,816	9448	9448
English	693,122	8655	9929
Gujarati	183,190	3028	2327
Hindi	531,513	3301	1819
Kannada	165,493	9760	8206
Marathi	312,808	9760	8206
Tamil	245,861	3885	4341
Telugu	274,927	3766	3474
All Languages	3,150,900	61562	57041

3.3 Proposed architecture

Figure 1 outlines the proposed methodology for building a language-independent dialogue summarisation and medical QA system.

3.4 Question and Answer architecture

The retrieval system uses a knowledge base consisting of: Medical literature & health forums (10; 23), Condition-specific documents, and translated multilingual knowledge bases (22; 30). Each document chunk is embedded and stored in the FAISS index for similarity-based retrieval (14). To strengthen factual grounding, a cross-

Figure 1: Block Diagram for the medical QA and dialogue summarisation

lingual fined-tuned RAG layer is integrated as shown in Figure 1 (16). This module retrieved the semantically relevant passages from a multilingual FAISS-based vector store before generating the final response to a user query. The retrieval setup included the Embeddings Model (intfloat/e5 model) (33), Vector Index (FAISS Flat Index), Similarity Metric (Cosine Similarity), and Top-K Retrieved Passages (5). The retrieved chunks are concatenated with the input query to create an enriched prompt, then processed through fined-tuned Sarvam 3B Model on QnA to generate the answer based on it (32).

3.5 Summarisation Architecture

As shown in Figure 1, the fine-tuned small-mT5 model is employed for the summarisation of the dialogues (27). The mT5 model generalises well across languages while maintaining contextual coherence. Key implementation details include a shared embedding layer for multilingual adaptability, 12 encoder and 12 decoder layers equipped with multi-head attention, a sequence-to-sequence design that encodes the dialogue and autoregressively decodes a summary, and the use of the prefix "summarise:" before each dialogue to guide task conditioning (21). The generated outputs are compared with the retrieved context using semantic similarity scoring (11) to ensure content alignment, language consistency checks, and redundancy removal (2; 19).

4 Experimental Setup

Table 3 delineates the parameters used for developing a model for dialogue summarisation and the QA task, along with the evaluation metrics used.

Table 3: Training configurations of the proposed RAG architecture for QA and fine-tuned mT5 model for dialogue summarisation.

Category + Parameter	Summarisation Model	RAG Model
Task	Summarisation	QnA Answering
Generation Model	mT5 (Multilingual T5)	Sarvam 3B
Embedding Model	T5 Embeddings	intfloat/e5-large
Base Model	google/mt5-small	Sarvam 3B
Pretraining Corpus	3.15M Dialogues	176k QA pairs
Supported Languages	101+ Languages	10 Indian Languages
Hardware	A100 GPU (Google Colab)	A100 GPU (Google Colab)
Optimizer	AdaFactor	AdamW
Learning Rate	Inverse square root decay	2.0e–05
Epochs/Steps	Many (pre-training)	1 (baseline, extendable)
Tokenizer	SentencePiece	SentencePiece
Training Objective	Span Corruption	Sequence-to-sequence
Pretraining Framework	TensorFlow + T5X	Hugging Face Transformers
Loss Function	Cross-Entropy	Cross-Entropy
Metrics	ROUGE, BLEU, BERTScore	Exact Match (EM), F1 Score

5 Results and Discussion

5.1 Results

This section presents the evaluation results and in-depth analysis of the Team Samvad multilingual system developed for the NLP4Health Shared Task (1) on Multilingual Health Dialogue Summarisation and Question Answering (23; 10). Our analysis spans QA, Summarisation (Text) across nine languages: Hindi, Bangla, Tamil, Telugu, Kannada, Gujarati, Marathi, Assamese, and English.

5.2 Discussion

As shown in figure 2, high F1 scores in Bangla (0.7803), Hindi (0.7479) and Marathi (0.39) reflect strong retrieval and semantic understanding of the proposed system (3), and low F1 scores in

Table 4: Test results on the QA task

Language	f1	bertscore_f1	cometscore
Marathi	0.3995	0.8392	0.3593
Kannada	0.2469	0.8375	0.4287
Gujarati	0.4235	0.8435	0.5626
English	0.2947	0.7960	0.5725
Telugu	0.2553	0.8416	0.4582
Tamil	0.2970	0.8299	0.4789
Bangla	0.7803	0.8144	0.4915
Hindi	0.7479	0.8376	0.4839
Assamese	0.4847	0.8055	0.4596

Table 5: Test results on the dialogue summarisation task

Language	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1	BERTScore F1	COMETScore
Marathi	0.2077	0.0597	0.1458	0.7766	0.4566
Kannada	0.1737	0.0511	0.1196	0.7875	0.4771
Gujarati	0.1664	0.0553	0.1170	0.7861	0.4586
English	0.1538	0.0535	0.1023	0.7952	0.4693
Telugu	0.1768	0.0593	0.1208	0.7934	0.4705
Tamil	0.1952	0.0574	0.1351	0.7927	0.4833
Bangla	0.2074	0.0588	0.1415	0.7824	0.4844
Hindi	0.1877	0.0518	0.1218	0.7939	0.4933
Assamese	0.1814	0.0576	0.1250	0.7939	0.4678

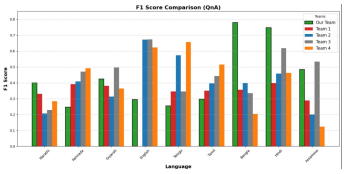

Figure 2: F1 score comparison for Q&A task on medical dialogue data.

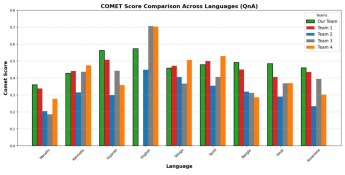

Figure 3: COMET score comparison for Q&A task on medical dialogue data.

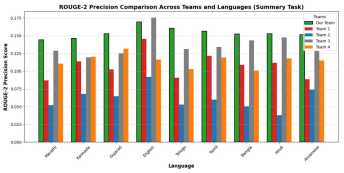

Figure 4: ROUGE-2 Precision comparison for summarisation task.

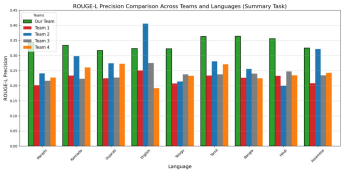

Figure 5: ROUGE-L Precision comparison for summarisation task.

Table 6: BERT F1 Scores comparison for the dialogue summarisation task

Language	Highest BERT F1	Our team BERT F1
Marathi	0.811	0.7766
Kannada	0.8247	0.7875
Gujarati	0.831	0.7861
English	0.8353	0.7952
Telugu	0.8344	0.7934
Tamil	0.8382	0.7927
Bangla	0.822	0.7824
Hindi	0.8364	0.7939
Assamese	0.8341	0.7939

Tamil, Telugu, and Kannada (<0.30). However, cometscore for the QA task in the Dravidian language was high, as shown in figure 3 (20). For the summarisation task as shown in figure 4, 5, the proposed model achieved the highest ROGUE2 and ROGUEL precision scores in all Indian languages (2). Comparable BERTScore values (0.77–0.79) to the highest results as shown in Table 6 indicate the model produces meaning-preserving paraphrases suitable for patient communication (19; 13).

Manual inspection of the QA and summarisation outputs revealed that the proposed model consistently preserved medical intent even when surface wording differed, often employing paraphrasing (e.g., “blocked nose” for “nasal obstruction”) (11). Responses were sometimes over-generated, providing explanatory answers rather than strictly extractive ones (16). Overall, the system delivers accurate patient-centered answers in multiple languages (10; 23). Limitations include lower performance in Dravidian languages, a need for structured generation in summary KnV outputs, and potential gaps in cultural or idiomatic understanding.

6 Conclusion

The results verified that integrating RAG with fine-tuned pre-trained language models (16; 21) can enhance the semantic understanding of the medical data without developing NLP systems from scratch. The proposed models achieved promising results as reflected in high BERTScore, F1 score, COMETScore and ROUGE precision (19; 3; 20; 2). The RAG architecture proved effective across all indic languages. (25). The Dravidian languages, such as Tamil, Telugu, and Kannada, still require improvisation (24). Team Samvad demonstrates the feasibility of a multilingual RAG-based system for medical dialogue understanding (23; 10).

7 Future Work

Future work’s primary focus will be on enhancing performance for languages such as Dogri and Assamese, including code-mixed inputs (25). To address this, we plan to target fine-tuning on medical datasets (e.g., PubMedQA (12) and MIMIC-III (26)) to improve factual accuracy, lexical precision, and overall summarisation performance. Optimisation of RAG thresholds and prompt design will be explored to enhance both summary fluency and coherence (16; 21).

Acknowledgment

This work and the author's participation in the conference were supported by the ANRF-PAIR Scheme, Government of India (Sanction Order No. ANRF/PAIR/2025/000008/PAIR).

References

[1] NLP-AI4Health 2025 Shared Task Overview. *AACL-IJCNLP 2025 Workshop on NLP-AI4Health*. Available at: https://2025.nlpai4health.com/#shared-task

[2] C.-Y. Lin. *ROUGE: A Package for Automatic Evaluation of Summaries*. In *ACL Workshop on Text Summarization Branches Out*, 2004.

[3] M. Sokolova and G. Lapalme. *A Systematic Analysis of Performance Measures for Classification Tasks*. *Information Processing Management*, 45(4):427–437, 2009.

[4] M. Lui and T. Baldwin. *langid.py: An Off-the-shelf Language Identification Tool*. In *ACL System Demonstrations*, 2012.

[5] T. Kiss and J. Strunk. *Unsupervised Multilingual Sentence Boundary Detection*. *Computational Linguistics*, 32(4):485–525, 2006.

[6] R. Sproat and N. Jaitly. *RNN Approaches to Text Normalization: A Challenge*. In *Interspeech*, 2016.

[7] A. See, P. J. Liu, and C. D. Manning. *Get To The Point: Summarization with Pointer-Generator Networks*. In *ACL*, 2017.

[8] H. Lin, J. Zhu, and J. Zhang. *Global Encoding for Abstractive Summarization*. In *ACL*, 2018.

[9] E. Alsentzer, J. Murphy, W. Boag, W. H. Weng, D. Jin, T. Naumann, and M. McDermott. *Publicly Available Clinical BERT Embeddings*. In *Proceedings of the 2nd Clinical NLP Workshop*, 2019.

[10] A. Ben Abacha and D. Demner-Fushman. *MedQuAD: Medical Question Answering Dataset Containing Question-Answer Pairs from Trusted Medical Sources*. In *BioNLP Workshop and Shared Task*, 2019.

[11] N. Reimers and I. Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. In *EMNLP*, 2019.

[12] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu. *PubMedQA: A Dataset for Biomedical Research Question Answering*. In *EMNLP*, 2019.

[13] Q. Ma, J. Wei, O. Bojar, and Y. Graham. *Results of the WMT19 Metrics Shared Task*. In *WMT*, 2019.

[14] J. Johnson, M. Douze, and H. Jégou. *Billion-Scale Similarity Search with GPUs*. *IEEE Transactions on Big Data*, 2019.

[15] K. Huang, J. Altosaar, and R. Ranganath. *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission*. *arXiv:1904.05342*, 2019.

[16] P. Lewis, E. Pérez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, and S. Riedel. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. In *NeurIPS*, 2020.

[17] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation*. In *ACL*, 2020.

[18] J. Lee, W. Yoon, S. Kim, D. Kim, C. So, and J. Kang. *BioBERT: A Pre-trained Biomedical Language Representation Model*. *Bioinformatics*, 36(4):1234–1240, 2020.

[19] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. *BERTScore: Evaluating Text Generation with BERT*. In *ICLR*, 2020.

[20] R. Rei, C. Stewart, A. Farinha, and A. Lavie. *COMET: A Neural Framework for MT Evaluation*. In *EMNLP*, 2020.

[21] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[22] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. *Unsupervised Cross-lingual Representation Learning at Scale*. In *ACL*, 2020.

[23] Q. Chen, Z. Liu, K. Ding, and W. Wang. *MedDialog: Large-scale Medical Dialogue Datasets*. In *EMNLP*, 2020.

[24] M. Baskar, B. R. Chakravarthi, and S. Thavareesan. *DravidianCodeMix: Sentiment Analysis Dataset for Dravidian Languages in Code-Mixed Text*. In *ICON*, 2020.

[25] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. *The State and Fate of Linguistic Diversity and Inclusion in the NLP World*. In *ACL*, 2020.

[26] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. *MIMIC-III: A Freely Accessible Critical Care Database*. *Scientific Data*, 3:160035, 2016.

[27] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, S. Narang, M. Matena, Y. Zhou, S. Kale, and C. Raffel. *mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer*. In *NAACL*, 2021.

[28] T. Gao, X. Yao, and D. Chen. *SimCSE: Simple Contrastive Learning of Sentence Embeddings*. In *EMNLP*, 2021.

[29] Q. Lhoest, A. Villanova del Moral, Y. Jernite, A. Thakur, P. von Platen, et al. *Datasets: A Community Library for Natural Language Processing*. In *EMNLP System Demonstrations*, 2021.

[30] J. Tiedemann. *The Tatoeba Translation Challenge: Realistic Data Sets for Low-Resource and Multilingual MT*. In *WMT*, 2020.

[31] Y. Wang, Z. Wang, S. Shen, and Y. Yang. *Text Embeddings by Weakly-Supervised Contrastive Pre-training*. *arXiv:2302.08582*, 2023.

[32] S. Sharma, P. Kumar, and S. Agarwal. *Sarvam: Multilingual Open Large Language Models for India*. *arXiv:2403.05696*, 2024.

[33] L. Wang, C. Geng, X. Ma, H. Yang, and F. Wei, *Text Embeddings by Weakly-Supervised Contrastive Pre-training*. *arXiv preprint arXiv:2212.03533*, 2022.