Faithful Summarization of Consumer Health Queries: A Cross-Lingual Framework with LLMs

Ajwad Abrar *
Islamic University of Technology
Dhaka, Bangladesh
ajwadabrar@iut-dhaka.edu

Nafisa Tabassum Oeshy *
Islamic University of Technology
Dhaka, Bangladesh
nafisatabassum4@iut-dhaka.edu

Prianka Maheru *
Islamic University of Technology
Dhaka, Bangladesh
priankamaheru@iut-dhaka.edu

Farzana Tabassum
Islamic University of Technology
Dhaka, Bangladesh
farzana@iut-dhaka.edu

Tareque Mohmud Chowdhury
Islamic University of Technology
Dhaka, Bangladesh
tareque@iut-dhaka.edu

Abstract

Summarizing consumer health questions (CHQs) can ease communication in healthcare, but unfaithful summaries that misrepresent medical details pose serious risks. We propose a framework that combines TextRank-based sentence extraction and medical named entity recognition with large language models (LLMs) to enhance faithfulness in medical text summarization. In our experiments, we fine-tuned the LLaMA-2-7B model on the MeQSum (English) and BanglaCHQ-Summ (Bangla) datasets, achieving consistent improvements across quality (ROUGE, BERTScore, readability) and faithfulness (SummaC, AlignScore) metrics, and outperforming zero-shot baselines and prior systems. Human evaluation further shows that over 80% of generated summaries preserve critical medical information. These results highlight faithfulness as an essential dimension for reliable medical summarization and demonstrate the potential of our approach for safer deployment of LLMs in healthcare contexts.

1 Introduction

The rapid growth of online health consultations, especially consumer health questions (CHQs), has created both opportunities and challenges for healthcare delivery. These platforms, accelerated by the pandemic, now serve as vital sources of medical information and support. However, the large volume of verbose and sometimes redundant patient queries places a significant burden on healthcare professionals, who must spend time identifying the core concern before providing an appropriate response. Automatic medical text summarization has emerged as a potential solution to this problem by condensing lengthy questions into concise, focused forms.

Traditional approaches to text summarization are primarily evaluated on *general quality* metrics such as ROUGE or BERTScore, which measure lexical or semantic similarity to reference summaries. While useful, these metrics often fail to capture *faithfulness*—the factual consistency of a summary

*These authors contributed equally to the work.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Muslims in ML.

with its source. Prior studies show that abstractive models frequently produce intrinsic errors (e.g., misrepresenting entities or relationships) and extrinsic errors (e.g., introducing unsupported facts) [Maynez et al., 2020, Huang et al., 2021]. In the medical domain, even minor distortions can mislead professionals or patients, posing direct risks to health outcomes. Thus, ensuring faithful summaries is essential for practical deployment in healthcare contexts.

Despite the progress of large language models (LLMs), current systems still struggle to balance fluency, conciseness, and factual consistency in medical summarization. Faithfulness remains underexplored compared to readability or general accuracy, and most methods do not explicitly address it. This gap motivates the need for specialized frameworks that preserve the integrity of medical information while still providing concise and accessible summaries.

To address this, we propose a novel framework that combines **TextRank-based sentence extraction** with **medical named entity recognition (NER)** to guide LLMs in generating summaries that are both accurate and faithful. We evaluate the framework on English (MeQSum) and Bangla (BanglaCHQ-Summ) datasets using LLaMA-2-7B fine-tuned with low-rank adaptation. Our main contributions are:

- We integrate extractive and abstractive methods to improve both informativeness and reliability.
- We incorporate medical NER to ensure critical entities are preserved in summaries.
- We provide the first cross-lingual evaluation of faithfulness in medical CHQ summarization, demonstrating improvements over zero-shot baselines and prior state-of-the-art systems.

2 Related Work

Text summarization is a well-established task in NLP that aims to condense lengthy documents while preserving salient content [El-Kassas et al., 2021]. Techniques generally fall into extractive methods, which select important sentences, and abstractive methods, which generate new phrasing [Rush et al., 2017]. While both have advanced with neural models and pretrained transformers, applying them to biomedical text is especially challenging due to domain-specific terminology, the complexity of clinical narratives, and the high stakes of factual accuracy [Afantenos et al., 2005, Morozovskii and Ramanna, 2023].

Medical summarization has been studied for clinical notes, conversations, and consumer health questions (CHQs). The MeQSum dataset [Ben Abacha and Demner-Fushman, 2019] introduced 1,000 annotated CHQs and spurred research into neural abstractive approaches. More recently, BanglaCHQ-Summ [Khan et al., 2023] extended this task to Bangla, highlighting cross-lingual gaps. Despite promising results with transformer-based models [Michalopoulos et al., 2022], maintaining reliability across languages and medical domains remains an open challenge.

A key limitation of existing work lies in *faithfulness*—the factual consistency of generated summaries. Prior studies show abstractive models often introduce intrinsic errors (contradictions) or extrinsic errors (unsupported additions) [Maynez et al., 2020, Huang et al., 2021]. Recent solutions include constrained decoding [Mao et al., 2020], fact-checking modules [Kryscinski et al., 2020], and specialized evaluation metrics such as SummaC and AlignScore [Laban et al., 2021, Zha et al., 2023]. However, ensuring factual alignment for consumer health queries, where even minor distortions can mislead patients or clinicians, remains underexplored. This gap motivates our focus on frameworks that explicitly preserve medical entities and source-grounded content.

3 Proposed Methodology

We propose a framework to improve the faithfulness of medical text summarization, with a focus on Consumer Health Questions (CHQs) in English and Bangla. The framework combines *Medical Named Entity Recognition (NER)* and the *TextRank* algorithm for extractive sentence selection, followed by fine-tuning a Large Language Model (LLM) to generate accurate and reliable summaries. An overview is shown in Figure 1.

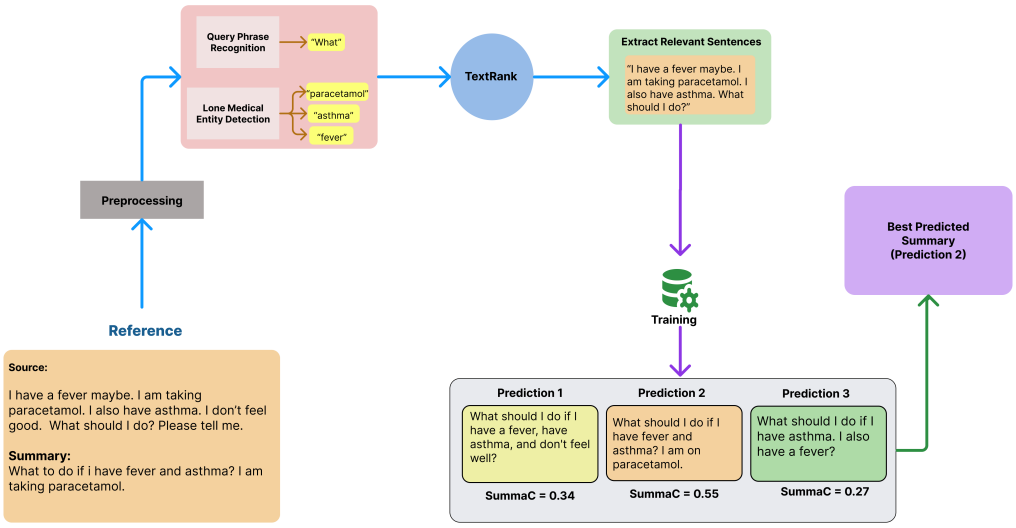

Figure 1: Proposed framework: TextRank extracts relevant sentences containing medical entities, which are used to fine-tune the LLM. The final summary is selected to maximize both accuracy and faithfulness.

3.1 Datasets

We use two benchmark datasets. For English, the **MeQSum** dataset [Ben Abacha and Demner-Fushman, 2019] contains 1,000 consumer health questions with expert-validated summaries, ensuring high-quality references. For Bangla, we use **BanglaCHQ-Summ** [Khan et al., 2023], which includes 2,350 annotated pairs of questions and summaries, representing the first large-scale resource for Bangla CHQ summarization. Together, these datasets enable evaluation across both high-resource and low-resource settings.

3.2 Preprocessing and Relevant Sentence Extraction

Preprocessing standardizes the datasets into *question* and *summary* fields, while identifying overlapping medical entities and negation terms to ensure critical information is retained. To reduce noise, we apply the **TextRank** algorithm [Mihalcea and Tarau, 2004] to extract sentences containing medical entities and query-related words. This guarantees that summaries remain faithful to medically important content before abstractive generation by the LLM.

3.3 Evaluation Metrics

We assess performance using both general quality and faithfulness metrics. General quality is measured by ROUGE-1/2/L and BERTScore, which capture lexical and semantic overlap. Faithfulness is measured by SummaC [Laban et al., 2021] and AlignScore [Zha et al., 2023], which evaluate factual consistency, along with Flesch Reading Ease (FRE) for readability. This combination ensures that generated summaries are not only fluent but also factually aligned with the source.

4 Results and Discussion

4.1 Performance Analysis (MeQSum)

We evaluate LLaMA-2-7B in four settings: zero-shot, fine-tuning (FT) without TextRank, FT with TextRank-selected sentences (FT+TR), and best-of-3 selection using either ROUGE-1 (R1) or SummaC as the selector. Zero-shot underperforms, as expected for task-specific summarization

Setting (MeQSum)	R1	R2	RL	BERT	Read.	SummaC	AlignScore
Zero-shot (no FT)	21.97	6.48	19.98	0.60	65.16	0.28	21.80
FT (no TR)	44.23	27.36	41.55	0.71	70.21	0.31	38.45
FT + TR	47.07	29.44	44.08	0.72	70.69	0.37	45.65
Best-of-3 (R1 select)	**50.50**	**34.38**	**47.74**	**0.74**	**71.56**	0.40	39.24
Best-of-3 (SummaC)	48.27	31.38	45.34	0.73	71.56	**0.57**	**45.91**

Table 1: MeQSum results across settings. Best-of-3 improves both quality and faithfulness; SummaC-based selection yields the highest factual consistency and strong alignment.

Model	R1	R2	RL	BERT	Read.	SummaC	AlignScore
Mixtral-8x7B-Inst. [Dada et al., 2024]	32.47	36.38	16.86	0.72	–	–	–
BioBART + FaMeSumm [Zhang et al., 2023]	31.76	11.71	29.64	0.74	–	0.46	–
Ours (Best-of-3)	**50.50**	**34.38**	**47.74**	**0.74**	**71.56**	**0.57**	**0.46**

Table 2: State-of-the-art comparison on MeQSum. Our method achieves the best performance across most metrics, with strong factual consistency (SummaC) and alignment (AlignScore).

[Abbasiantaeb et al., 2024]. Fine-tuning substantially boosts both general quality and faithfulness, and adding TextRank further improves alignment with source content. Selecting the best of three candidates yields the strongest scores.

Temperature. We sweep $t \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and observe a trade-off: lower t favors ROUGE, higher t favors SummaC. We adopt t=0.7 as a balanced choice, giving peak faithfulness with competitive ROUGE [Yu et al.].

Comparison to prior work. Our approach surpasses strong baselines on MeQSum, particularly in R1/RL and readability, while achieving competitive or superior faithfulness.

4.2 Performance Analysis (BanglaCHQ-Summ)

We replicate the evaluation on BanglaCHQ-Summ to assess cross-lingual robustness. Zero-shot performance is weak, reflecting linguistic/domain gaps. Fine-tuning improves all metrics, and FT+TR yields further gains. Best-of-3 selection again provides the strongest outcomes; SummaC-based selection maximizes faithfulness.

Setting (Bangla)	R1	R2	RL	BERT	SummaC
Zero-shot (no FT)	19.10	8.21	18.97	0.62	0.22
FT (no TR)	28.24	14.22	24.54	0.71	0.26
FT + TR	30.71	15.71	28.95	0.74	0.28
Best-of-3 (R1 select)	**32.35**	**16.32**	**29.09**	**0.76**	0.29
Best-of-3 (SummaC select)	30.92	15.74	27.35	0.73	**0.32**

Table 3: BanglaCHQ-Summ results. Best-of-3 improves quality (R1/RL/BERT) and faithfulness (SummaC).

4.3 Human Evaluation

To validate the reliability of generated summaries, we conducted a human evaluation on the MeQSum dataset with the help of a medical doctor. The evaluation focused on two questions: (1) whether the summary retained all critical information from the source text, and (2) whether it was factually consistent. A binary (yes/no) judgment was provided for each case. A summary was considered faithful only if both conditions were met. Results showed that 82% of the summaries satisfied these criteria, demonstrating strong factual alignment.

5 Conclusion

We proposed a framework that combines TextRank-based extraction, medical NER, and fine-tuned LLaMA-2-7B to enhance the faithfulness of medical text summarization. Experiments on English (MeQSum) and Bangla (BanglaCHQ-Summ) datasets show consistent gains over zero-shot and prior systems in both quality and factual consistency. Human evaluation further confirmed the reliability of our approach, with over 80% of summaries judged as faithful. A limitation of this study is that experiments were conducted on a single LLM (LLaMA-2-7B) and two languages, leaving scope for future work to explore multiple LLMs and broader multilingual settings. Future directions include adapting the framework to few-shot and multilingual settings, extending to other sensitive domains (e.g., legal and financial texts), and incorporating expert feedback for improved robustness and practical integration into clinical workflows.

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 8–17, 2024.

Stergos Afantenos, Vangelis Karkaletsis, and Panagiotis Stamatopoulos. Summarization from medical documents: a survey. *Artificial intelligence in medicine*, 33(2):157–177, 2005.

Asma Ben Abacha and Dina Demner-Fushman. On the summarization of consumer health questions. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1215. URL https://aclanthology.org/P19-1215.

Amin Dada, Marie Bauer, Amanda Butler Contreras, Osman Alperen Koraş, Constantin Marc Seibold, Kaleb E Smith, and Jens Kleesiek. Clue: A clinical language understanding evaluation for llms. *arXiv preprint arXiv:2404.04067*, 2024.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679, 2021.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*, 2021.

Alvi Khan, Fida Kamal, Mohammad Abrar Chowdhury, Tasnim Ahmed, Md Tahmid Rahman Laskar, and Sabbir Ahmed. BanglaCHQ-summ: An abstractive summarization dataset for medical queries in Bangla conversational speech. In Firoj Alam, Sudipta Kar, Shammur Absar Chowdhury, Farig Sadeque, and Ruhul Amin, editors, *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 85–93, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.banglalp-1.10. URL https://aclanthology.org/2023.banglalp-1.10.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.750. URL https://aclanthology.org/2020.emnlp-main.750.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. Summac: Re-visiting nli-based models for inconsistency detection in summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8425–8441, 2021.

Yuning Mao, Xiang Ren, Heng Ji, and Jiawei Han. Constrained abstractive summarization: Preserving factual consistency with constrained generation. *arXiv preprint arXiv:2010.12723*, 2020.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.349. URL https://aclanthology.org/2022.findings-emnlp.349.

Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

Danila Morozovskii and Sheela Ramanna. Rare words in text summarization. *Natural Language Processing Journal*, 3:100014, 2023.

AM Rush, SEAS Harvard, S Chopra, and J Weston. A neural attention model for sentence summarization. aclweb. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2017.

Chan Xing Yu, Chan Si Yu James, and Poh Hui-Li Phyllis David. Can llms have a fever? investigating the effects of temperature on llm security.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*, 2023.

Nan Zhang, Yusen Zhang, Wu Guo, Prasenjit Mitra, and Rui Zhang. Famesumm: Investigating and improving faithfulness of medical summarization, 2023.