UTSamuel at ArchEHR-QA 2025: A Clinical Question Answering System for Responding to Patient Portal Messages Using Generative AI

Samuel M. Reason, Liwei Wang, Hongfang Liu, Ming Huang*
McWilliams School of Biomedical Informatics,
University of Texas Health Science Center at Houston
***Corresponding author (Ming.Huang@uth.tmc.edu)**

Abstract

Responding to patient portal messages places a substantial burden on clinicians. To mitigate this, automatically generating answers to patient questions by considering their medical records is a critical solution. In this study, we proposed a clinical question answering system for the BioNLP 2025 Shared Task on Grounded Electronic Health Record Question Answering. The system processed each patient message case by selecting relevant sentences as evidences from the associated clinical notes and generating a concise, medically accurate answer to the patient's question. A generative AI model from OpenAI (GPT-4o) was leveraged to assist with sentence selection and answer generation. Each response is grounded in source text, limited to 75 words, and includes sentence-level citations. The system was evaluated on 100 test cases using alignment, citation, and summarization metrics. Our results indicate the significant potential of the clinical question answering system based on generative AI models to streamline communication between patients and healthcare providers by automatically generating responses to patient messages.

1 Introduction

Patient portal messaging has become a critical communication channel between patients and healthcare providers, extending interaction beyond scheduled visits (Huang, Fan et al. 2022, Huang, Khurana et al. 2023). This platform enables dynamic exchanges on complex issues like new symptoms, disease follow-ups, medication concerns, and other medical inquiries (De, Huang et al. 2021, Huang, Wen et al. 2022).

With the increasing adoption of digital technologies by healthcare organizations to foster patient engagement and care, patient portals have become more prevalent, leading to a substantial surge in portal message volume (Huang, Khurana et al. 2022, Zhou, Arriaga et al. 2022). While this increased communication holds the promise of improved patient care and satisfaction, it has also created challenges in terms of efficient management and timely responses. Consequently, secure messaging has contributed to a heavier workload and burnout among clinicians by increasing patient-clinician interactions between in-person visits. For instance, primary care physicians commonly spend 1.5 hours daily processing around 150 inbox messages, often extending their work beyond regular clinic hours (Akbar, Mark et al. 2021). This constant influx of patient messages has become a significant stressor in clinical settings, particularly for primary care physicians, exacerbating burnout. Thus, the development of a clinical question answering system that can automatically generate answers to patient questions derived from their messages is essential to aid clinicians in responding effectively to patient portal communications (Ren, Wu et al. 2023, Ren, Wu et al. 2024).

The BioNLP 2025 shared task on grounded question answering (QA) from electronic health records (EHRs) focuses on automatically generating answers to patients' health-related questions that are grounded in the evidence from patients' clinical notes (Soni and Demner-Fushman 2025a). This QA task emphasizes direct citation of supporting evidence and grounding within the relevant clinical notes of patients. The need for accurate, transparent, and reproducible QA methods is especially important in clinical settings, where misinterpretation or hallucination can lead to critical errors.

This paper presents a clinical QA system developed leveraging generative AI models. The system selects sentences relevant to the clinical

BioNLP 2025 Shared Tasks, pages 91–95

question and uses them to generate a plain-language response. No training data, external models, or automation was used. The emphasis throughout development was on traceability, consistency, and alignment with the shared task format.

2 Methods

2.1 Dataset

The dataset for this task includes patient questions (based on real patient queries) and associated EHR data (from MIMIC-III) containing vital clinical evidence (Soni and Demner-Fushman 2025b). Each question-note combination is a "case." Clinical note excerpts are provided with pre-assigned sentence numbers, which systems must use for citing evidence. Additionally, each sentence is manually annotated with a "relevance" label ("essential," "supplementary," or "not-relevant") indicating its role in answering the question. The development set of 20 cases provides these relevance labels. The test set contains 100 cases without the relevance labels.

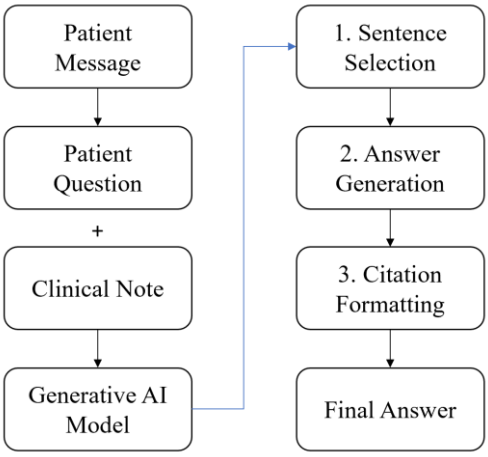

Figure 1: Overview of the clinical QA system for the BioNLP 2025 shared task

2.2 System Design

As shown in Figure 1, the clinical QA system for automatically generating answers to patient questions was implemented as a three-step pipeline applied for each patient message:

1. Sentence Selection – identifying essential and supplementary sentences from the clinical note
2. Answer Generation – using a structured prompt to compose a response with Generative AI models
3. Citation Formatting – ensuring each sentence is properly cited using its unique sentence ID

All work was done directly in an interactive session of ChatGPT (GPT-4o) (Hurst, Lerer et al. 2024) through HIPAA compliant Azure OpenAI Studio, without the use of application programming interface (APIs) and model fine-tuning.

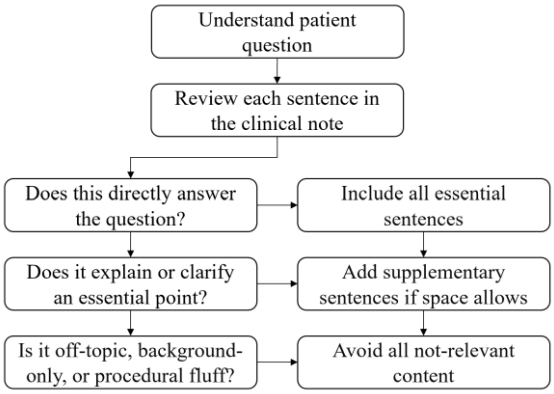

Figure 2: Sentence selection logic

2.3 Sentence Selection Strategy

Each sentence from the clinical note excerpt was reviewed and categorized as one of the following three categories:

1. Essential – directly answers the clinician's question (e.g., diagnosis, treatment, hospital course)
2. Supplementary – adds clinical context (e.g., medications, labs, background)
3. Not relevant – unrelated or duplicative information

Sentences were selected based on clinical reasoning and their alignment with the question's intent. If essential information was distributed across multiple sentences or incomplete without context, supplementary sentences were added for clarity.

To refine the sentence selection process, we used the 20 development cases provided with gold-standard relevance annotations. These cases included sentence-level labels (essential, supplementary, or not relevant), which allowed us to evaluate how well different selection strategies automatically designed by generative AI aligned with human annotations. Several iterations of sentence triage logic were tested and revised based on comparisons to these keys. The final sentence selection logic is illustrated in Figure 2. The process to develop the final sentence selection

logic is detailed in Section 2.3.1. Once finalized, this selection logic was applied to the 100 test cases, which were processed without access to gold relevance labels.

2.3.1 Self-evaluation Approach for Sentence Selection

To improve sentence selection, we used ChatGPT to perform a self-evaluation analysis based on the 20-case development set. After ChatGPT generated answers without access to these labels, we uploaded the annotations to assess its performance by providing the following prompt:

"I'm going to upload a file which, for each case, shows which sentences are essential, supplementary, or not relevant. I want you to analyze how you did on using essential sentences for your answer—how many did you use, how many did you miss, etc.? Do the same for supplementary and not-relevant sentences."

Based on this self-evaluation analysis, ChatGPT recommended several logic changes, which we adopted in the final system: (1) Read the clinician question to determine its clinical focus (e.g., diagnosis, treatment, prognosis). (2) Classify note sentences as essential, supplementary, or not relevant based on their alignment with the question. (3) Generate the answer by including essential sentences first, then supplementary ones if needed. This refinement process—enabled by prompting ChatGPT to self-assess—improved the completeness of generated answers, particularly in aligning with the information explicitly required by the question.

2.4 Answer Generation and Citation Formatting

The answer to the patient question was composed using a structured prompt. The prompt included the question, the selected sentences, and explicit instructions. A typical prompt was as follows:

"Write a medically accurate answer to the question below using only the sentences provided. Limit the answer to 75 words. Keep the language clear and professional. At the end of each sentence in your answer, cite the original sentence ID in this format: |ID|."

Additionally, the prompt strategy enables all generated answers to: (1) remain under 75 words (2) cite each supporting sentence using its ID (|sentence_id|) (3) use only content from the provided note excerpt (4) be written in medically appropriate, clear language. This constraint-based format ensured that responses were traceable and aligned with the evidence selection.

2.5 Evaluation

The generated answers will be evaluated on two key aspects: Factuality (how well they are grounded in clinical evidence) and Relevance (how well they answer the question). Factuality is measured using Precision, Recall, and F1 scores by comparing the evidence sentences cited in the generated answer against a manually annotated ground truth set of essential and supplementary sentences. Two F1 scores are calculated: a strict score considering only 'essential' sentences as correct evidence, and a lenient score including both 'essential' and 'supplementary' sentences. Relevance is assessed by comparing the generated answer text to the ground truth 'essential' sentences and the original question using metrics like BLEU (Papineni, Roukos et al. 2002), ROUGE (Lin 2004), SARI (Xu, Napoles et al. 2016), BERTScore (Zhang, Kishore et al. 2019), AlignScore (Zha, Yang et al. 2023), and MEDCON (Yim, Fu et al. 2023). The overall score for ranking will be the average of the strict Factuality F1 score

Metric		Min	Max	Mean	Median	Score
Overall	Overall	19.3	53.7	39.8	39.2	37.8
	Factuality	13.2	60.5	47.7	45.3	47.8
	Relevance	25.2	48.8	31.8	33.1	27.8
Factuality	Strict F1(i)	13.2	60.5	47.7	45.3	47.8
	Strict F1(a)	18.7	62.6	51.4	48.5	49.0
	Lenient F1(i)	13.5	62.7	48.8	46.4	49.7
	Lenient F1(a)	18.6	64.8	52.6	50.0	51.8
Relevance	BLEU	0.1	14.3	1.7	2.6	0.6
	ROUGE	15.2	46.5	22.7	24.3	20.0
	SARI	36.7	73.1	54.4	55.5	56.7
	BERTScore	19.9	53.9	26.3	28.3	24.2
	AlignScore	35.2	92.4	52.9	54.2	35.4
	MEDCON	23.2	49.3	32.9	33.8	29.6

*F1(i) and F(a) denote F1 (micro) and F1 (macro), respectively.

Table 1: Analysis of key performance metrics

and a combined score derived from the normalized Relevance metrics.

3 Results

The clinical QA system was evaluated on 100 test cases using the official metrics provided by the shared task organizers. Its key performance metrics among the 30 participants are listed in Table 1.

Among the overall metrics (Factuality and Relevance), Factuality performance was relatively strong at 47.8, exceeding both the mean (47.7) and median (45.3). This indicates a consistent use of relevant evidence sentences. The strict and lenient micro F1 scores (47.8 and 49.7, respectively) were also higher than their respective means and medians.

For Relevance, the system scored 27.8, slightly lower than mean (31.8) and median (33.1). The score of SARI (56.7) is higher than mean (54.4) and median (55.5), suggesting the answers were readable and cleanly edited. However, metrics like ROUGE-Lsum (20.0), BLEU (0.6), BERTScore (24.2), and MEDCON (29.6) were slightly lower than mean and median because the system focused on giving short, evidence-backed answers rather than exact matches to the reference summaries.

4 Discussion

This study explored using a single generative AI model (GPT-4o) through OpenAI's interact session and prompts to generate answers to patient questions with evidence from their medical records. Our goal focused on the straightforward application of readily accessible generative AI models via the interact session, rather than developing complex clinical QA models. This approach leverages the easy deployment of generative AI, which bypasses the need for in-depth model development expertise such as API calls, fine-tuning, and knowledge injection.

The performance of the clinical QA system was comparable to the mean and median, indicating the feasibility of using a single generative AI for answering patient questions via direct interaction. Its stronger performance in Factuality compared to the mean and median highlights the effectiveness of the designed sentence selection logic in consistently utilizing relevant evidence for answer generation.

5 Limitations

The system was developed under tight time constraints. While the current system only used a single generative AI model and straightforward interactive workflow, our plans included experimentation with multiple strategies involving different generative AI models for a hybrid system, collaborative learning, and advanced evidence sentence classification. These extensions were not explored due to lack of time.

The generative AI model (ChatGPT) was accessed through the web interface for simplicity. Although the interactive session allows the ease use of generative AI models, this limited reproducibility and scalability. The interactive nature of the workflow made it difficult to test multiple prompting strategies at scale or implement programmatic validation. Use of the API could have enabled more consistent experimentation and versioning.

6 Conclusion

We present a clinical QA system developed through an interactive workflow with a generative AI model. The system selects relevant sentences and uses them to construct a short, evidence-grounded answer with sentence-level citations. No model fine-tuning or APIs were required. Our findings show the feasibility of the strategy to develop a clinical QA system for generating answers to patient questions in portal messages. Future work may explore multi-model workflows, collaborative learning, and more structured evaluation pipelines.

Acknowledgments

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under award numbers R01LM011934, the National Human Genome Research Institute under award number R01HG012748, the National Institute of Aging under award number R01AG072799, and the Cancer Prevention Institute of Texas (CPRIT) under award number RR230020. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine, the National Human Genome Research Institute, the National Institutes of Health, or the State of Texas.

References

Akbar, F., G. Mark, E. M. Warton, M. E. Reed, S. Prausnitz, J. A. East, M. F. Moeller and T. A. Lieu (2021). "Physicians' electronic inbox work patterns and factors associated with high inbox work duration." Journal of the American Medical Informatics Association 28(5): 923-930.

De, A., M. Huang, T. Feng, X. Yue and L. Yao (2021). "Analyzing patient secure messages using a fast health care interoperability resources (FIHR)–based data model: development and topic modeling study." Journal of medical Internet research 23(7): e26770.

Huang, M., J. Fan, J. Prigge, N. D. Shah, B. A. Costello and L. Yao (2022). "Characterizing patient-clinician communication in secure medical messages: retrospective study." Journal of Medical Internet Research 24(1): e17273.

Huang, M., A. Khurana, G. Mastorakos, A. Wen, H. He, L. Wang, S. Liu, Y. Wang, N. Zong and J. Prigge (2022). "Patient portal messaging for asynchronous virtual care during the COVID-19 pandemic: retrospective analysis." JMIR Human Factors 9(2): e35187.

Huang, M., A. Khurana, G. Mastorakos, J. Zhou, N. Zong, Y. Yu, J. E. Prigge, C. A. Patten, H. Liu and B. A. Costello (2023). Characterizing the Users of Patient Portal Messaging: A Single Institutional Cohort Study. 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI).

Huang, M., A. Wen, H. He, L. Wang, S. Liu, Y. Wang, N. Zong, Y. Yu, J. E. Prigge and B. A. Costello (2022). "Midwest rural‐urban disparities in use of patient online services for COVID‐19." The Journal of Rural Health 38(4): 908-915.

Hurst, A., A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes and A. Radford (2024). "Gpt-4o system card." arXiv preprint arXiv:2410.21276.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. Text summarization branches out.

Papineni, K., S. Roukos, T. Ward and W.-J. Zhu (2002). Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting of the Association for Computational Linguistics.

Ren, Y., D. Wu, A. Khurana, G. Mastorakos, S. Fu, N. Zong, J. Fan, H. Liu and M. Huang (2023). Classification of Patient Portal Messages with BERT-based Language Models. 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI).

Ren, Y., Y. Wu, J. W. Fan, A. Khurana, S. Fu, D. Wu, H. Liu and M. Huang (2024). "Automatic uncovering of patient primary concerns in portal messages using a fusion framework of pretrained language models." Journal of the American Medical Informatics Association 31(8): 1714-1724.

Soni, S. and D. Demner-Fushman (2025a). Overview of the ArchEHR-QA 2025 Shared Task on Grounded Question Answering from Electronic Health Records. The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. Vienna, Austria, Association for Computational Linguistics.

Soni, S. and D. Demner-Fushman (2025b). "A Dataset for Addressing Patient's Information Needs related to Clinical Course of Hospitalization." arXiv preprint.

Xu, W., C. Napoles, E. Pavlick, Q. Chen and C. Callison-Burch (2016). "Optimizing statistical machine translation for text simplification." Transactions of the Association for Computational Linguistics 4: 401-415.

Yim, W.-w., Y. Fu, A. Ben Abacha, N. Snider, T. Lin and M. Yetisgen (2023). "Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation." Scientific data 10(1): 586.

Zha, Y., Y. Yang, R. Li and Z. Hu (2023). "AlignScore: Evaluating factual consistency with a unified alignment function." arXiv preprint arXiv:2305.16739.

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi (2019). "Bertscore: Evaluating text generation with bert." arXiv preprint arXiv:1904.09675.

Zhou, J., R. I. Arriaga, H. Liu and M. Huang (2022). A Tale of Two Perspectives: Harvesting System Views and User Views to Understand Patient Portal Engagement. 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI).