BreastCancerDiagNet - Transformer-Based Clinical Question Generation for Automated History Taking

Maleeha Fathima
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India
maleeha.222@gmail.com (corresponding author)

Moulana Mohammed
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India
moulana@kluniversity.in

Received: 20 September 2025 | Revised: 11 October 2025 and 23 October 2025 | Accepted: 24 October 2025

 DOI: https://doi.org/10.48084/etasr.14966

ABSTRACT

History-taking is a highly significant procedure in clinical decision-making that remains time-consuming, inconsistent, and can lead to omissions. This study presents BreastCancerDiagNet, a complex transformer-powered system to automate medical history-taking and aid in diagnosis. This model uses structured patient demographics and unstructured clinical symptoms with hybridized ClinicalBERT embeddings, BiLSTM sequence modeling, and self-attention mechanisms. These capabilities are integrated in an encoder-decoder architecture with rotary position embeddings and FlashAttention to enable long-sequence processing. A Reinforcement Learning with Human Feedback (RLHF) strategy is used to refine the question generation strategy to reflect contextual reference to clinical practice. The proposed system was trained and tested using a breast cancer dataset of curated demographic data, symptom data, comorbidities, lifestyle indicators, and physician-curated ground truth questionnaires. The results show that BreastCancerDiagNet achieved a BLEU-4 score of 0.42, a ROUGE-L score of 0.56, and a BERT-F1 score of 0.88, which are higher than the Seq2Seq and Vanilla Transformer baselines. Qualitative analysis confirmed the relevance of the questions generated in clinical practice, covering lump presence, pain, discharge, family history, and drug use. The findings demonstrate the possibility of using BreastCancerDiagNet to save time in consultations, reduce the number of diagnostic errors, and serve as a future-generation Clinical Decision Support System (CDSS) that can be scaled and interpreted.

Keywords-medical question generation; transformer; clinicalBERT; reinforcement learning with human feedback; breast cancer diagnosis; clinical decision support

I. INTRODUCTION

Medical history-taking is the basis for diagnostic reasoning and clinical decision-making, where specialists can recognize the uniqueness of patient conditions and design the strategy to manage them in the most efficient manner possible. However, traditional methods in history-taking remain time-consuming, often incomplete, and dependent on clinical judgment that can cause a partial history and delay in diagnosis [1]. As the volume of patient information increases exponentially and the complexity of healthcare systems increases at an alarming pace, there are immediate calls for automated and intelligent frameworks that can help simplify the process without sacrificing its accuracy or interpretability.

The emergence of Artificial Intelligence (AI) and Large Language Models (LLMs) has radically changed the healthcare industry, offering new opportunities in terms of Electronic Health Records (EHRs), clinical documentation, and diagnostic support [2]. Unlike rule-based systems, recent transformer-based systems, such as GatorTronGPT and other domain-adapted LLMs, have shown the ability to create coherent and context-based clinical discourses and can perform knowledge-intensive medical activities [3]. Similarly, Question Answering (QA) and Question Generation (QG) systems have been introduced as promising technologies to ease the patient-provider communication process and help clinicians gather structured and unstructured medical knowledge [4].

In [5], a systematic review highlighted the idea that medical QA systems used at the point of care can improve efficiency by retrieving highly relevant information, easing the cognitive load on the physician. Retrieval-Augmented Generation (RAG)

models deployed in HER systems have led to a more accurate summarization and extraction of clinically critical information, on which downstream decisions are made [6]. Despite this, there are still issues, such as insufficient flexibility in the patient population, poor explainability, and the inability to integrate with clinical practice in the real world [7]. Transformer-based NLP applications, such as self-attention, domain-specific pretraining, and hybrid embedding planning schemes, have shown a major improvement in performance on biomedical datasets, clearing the divide between general-purpose language models and clinical-domain applications [8]. However, although generative AI in medicine has spread to a variety of applications, including diagnostic prediction, case summarization, and chatbots, there are still concerns about model reliability, control over hallucinations, and bias reduction [9].

Recent studies have shown that ensemble methods that capitalize on multiple LLMs can address some of these issues, achieving greater accuracy, greater robustness, and domain adaptation in clinical QA tasks [10]. However, not many models have applied these developments to the area of automated medical history-taking, where QG in a dynamic and personalized setting may significantly minimize diagnostic errors and speed up patient triage. The proposed system, called BreastCancerDiagNet, combines organized patient demographics with contextual embeddings of ClinicalBERT, BiLSTM, and self-attention. Enhanced with rotary position embeddings and FlashAttention, this system can produce adaptive and clinically relevant questions based on information about the specific patient. BreastCancerDiagNet is an interpretable, scalable, and efficient next-generation Clinical Decision Support System (CDSS) that uses Reinforcement Learning with Human Feedback (RLHF) and continuous retraining on clinician-curated reports.

A. *Objectives*

This study aimed to:

- Develop a transformer-based framework that automates medical history-taking by generating patient-specific and clinically relevant questions.
- Integrate structured patient demographics with advanced embeddings from ClinicalBERT, BiLSTM, and self-attention to produce context-rich representations for accurate question generation.
- Enhance the efficiency and scalability of the model using rotary position embeddings, FlashAttention, and RLHF.

II. LITERATURE SURVEY

Big data deep learning models show increasing potential in healthcare use cases such as documentation, support in triage, and generation of clinical dialogues [11]. Their results support the need to use domain-specific pretraining, which BreastCancerDiagNet utilizes through ClinicalBERT embeddings to generate medical questions based on context. Biomedically trained models enhance the performance of medical NLP [12], and this study employs clinical embeddings instead of general-purpose transformers.

In [13], different approaches to explainable clinical NLP (e.g., attention visualization, post-hoc explanations) were reviewed, showing the clinical importance of explainable models. These results drive the design decisions of BreastCancerDiagNet, which are self-attention and coefficient/attribution views, to improve the trust of clinicians in the generated questions. In [14], automated medical question answering, neural architectures, knowledge integration methods, and evaluation issues (domain shift, limited gold data) were explored. Hybrid strategies incorporate text comprehension with structured knowledge, which informs the hybrid nature of BreastCancerDiagNet for symptoms, demographics, and language embeddings. The study in [15] addressed the domain shift and the lack of clinical annotations. To alleviate this, BreastCancerDiagNet uses RLHF and multi-metric assessment to preserve semantic fidelity. ROUGE [16] and BLEU [17] are used to validate lexical matches, whereas BERTScore [18] offers semantic matching beyond a mere overlap of n-gram similarity.

Multiple-Choice Questions (MCQs) have become common in the health sciences as a method of knowledge testing, but the generation of quality items is time-intensive. The development of AI, specifically LLMs, can allow for fast and stable generation of MCQs. The study in [18] addressed advantages, disadvantages, and ethical implications, providing future research directions and guidelines for the use of AI in the development of MCQs. In [19], one of the first structured medical Question-Answer (QA) pair generation systems was provided, dealing with the lack of high-quality annotated data in the clinical field. This model combined medical entity extraction, template-driven question construction, and answer retrieval, leading to a significant improvement in the relevance and factual accuracy of generated QA pairs. This model, based on the use of domain-specific knowledge bases and supervised learning, showed better answer correctness, medical named-entity precision, and clinical relevance than generic generative models, setting a baseline standard in the creation of automated medical QA datasets and downstream activity, including CDSSs, medical triage systems, and diagnostic reasoning. In [20], a new difficulty-controllable QG framework was proposed, with the aim of influencing the generation procedure with stepwise rewriting. This strategy breaks down the generation of questions into a series of rewrites, giving fine control of the level of complexity, language structure, and the level of cognitive challenge. This model could generate questions of different degrees of difficulty and stepwise transformation cues, retaining the semantic aspects of the source content and being superior to earlier QG systems.

The AI field is quickly finding its way into the healthcare sphere, making it more efficient and accurate in its clinical processes and diagnosis. Transformer models have become useful for producing brief discharge overviews, enabling clinicians to do less work without missing important information [21]. AI-powered tools are used for the detection of oral cancer, improving accessibility in low-resource settings [22], and machine learning ensembles are effective in chronic obstructive pulmonary disease diagnosis, proving the potential of AI to assist with more complicated medical decision-making [23].

TABLE I. RESEARCH GAP ANALYSIS IN AUTOMATED MEDICAL QUESTION GENERATION AND CLINICAL NLP

Research gap analysis		
Study	**Method**	**Identified gaps**
[11]	Reviewed LLMs in healthcare, highlighting applications and limitations	Accuracy, safety, and ethical concerns remain unresolved; limited clinical customization.
[12] (BioBERT)	Pretrained biomedical embeddings improved NER, RE, and QA tasks	Focused only on unstructured biomedical corpora; lacks integration with patient demographics/symptoms
[13]	Surveyed explainable clinical NLP, emphasizing interpretability approaches	Few models embed interpretability directly into QG systems, limiting clinician trust.
[14]	Reviewed automated medical QA and neural architectures, recommending hybrid approaches	Lack of systems combining structured clinical data with text embeddings for QG.
[15]	Surveyed biomedical QA tasks and datasets, noting domain adaptation challenges	Persistent issues with data sparsity, variability, and generalizability.
[16] (ROUGE)	Introduced ROUGE as a content overlap evaluation metric	Lexical-only metric; fails to capture the semantic correctness of clinical questions.
[17] (BLEU)	Proposed BLEU as a surface-form precision measure for text generation.	Insufficient for clinical settings; lacks semantic depth and contextual relevance.
[18]	Introduced BERTScore for semantic evaluation and discussed LLMs for MCQ generation.	Focused mainly on MCQ generation; limited exploration of free-form clinical QG tasks.
[19]	Surveyed seq2seq and transformer-based QG, emphasizing attention and controllability.	Lack of domain-specific enhancements for medical QG; limited adaptability to patient-specific inputs.
[20]	Reviewed biomedical text mining, integrating structured signals with unstructured data.	Limited real-world deployment; integration strategies remain underexplored for adaptive history-taking.

III. METHODOLOGY

The proposed BreastCancerDiagNet framework aims to automate medical history-taking by combining raw patient symptoms and intelligent questionnaire generation with predictive modeling to formulate a full medical history report. In this work, medical history-taking refers to the wider structured task of collecting clinical information, and diagnostic QG is the automated generation of follow-up questions that depend on patient-specific input. The general system architecture in Figure 1 shows the sequential flow of data between patient input and the creation of the resulting combined report.

Its process begins with raw symptoms given by the patient, recorded in a structured or unstructured format. Interpretation of these symptoms is performed in a feature extraction and model loading module, where relevant attributes are represented and converted to machine-interpretable formats, ensuring that categorical and numerical data are standardized for later analysis. The features obtained are used to train the model in generating the questionnaire, where the system can continue to advance the quality and specificity of the questions produced. After preprocessing, the system loads a trained model to produce a rudimentary questionnaire that is specific to the symptoms the patient presents. The answers to this questionnaire enable the model to define basic symptoms that are the basis of clinical decision support. The framework estimates the possible medical field or disease sector in which the reported symptoms are related by using classical classifiers such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and a fuzzy inference system. The narrower follow-up questions can be generated based on this prediction. KNN, SVM, and fuzzy inference were chosen as classical classifiers to act as low-latency triage mechanisms and route clusters of symptoms before question generation because they offer fast inferences and do not obscure the decision boundaries of the initial stage of decision support like more complex black-box models do.

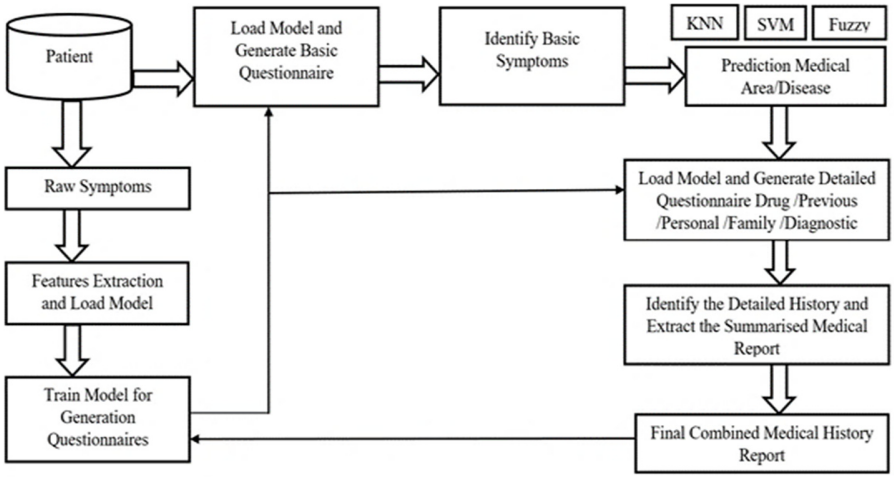

Fig. 1. System architecture of BreastCancerDiagNet.

The second step involves the creation of a comprehensive questionnaire that includes other areas such as drug history, past medical conditions, personal and family history, and diagnostic records. The system asks numerous questions to ensure that vital information is not neglected. The answer to these detailed questionnaires is used to determine the detailed medical history and derive a summarized report. The summarized report is included in the Final Combined Medical History Report, which incorporates patient demographics, symptom analysis, past medical history, and diagnostic indicators into an organized output. Clinicians can use this detailed report directly, saving time in consultation and reducing the number of diagnostic errors. In this way, the proposed method ensures a smooth process between patient input and clinical decision support and integrates machine learning classifiers, automatically generated questionnaires, and multistage history extraction.

Table II shows data samples collected from patients, representing raw information gathered directly during medical consultations. It provides a detailed snapshot of each patient's health status, serving as the foundation for further analysis and automated decision-making. In accordance with anonymization rules, the patient IDs and entries displayed in this table are artificial and are only used for demonstration purposes.

TABLE II. PATIENT RAW DATA

Attributes	**Patient Data**	
	Patient 1	**Patient 2**
Patient ID	1001	1002
Age	44	56
Married	Yes	Yes
Symptoms	patient came to hospital, complaint of lump 6 month back, associated pain in the right breast	history of lump in right breast since 1 month, no history of pain in the region
Present diseases	patient was complaining of lump at right breast since 6 months back and associated pain since 15 days back.	patient gives history of appearance of a small lump in right breast, which grew in size with time to attain the present size since 1 month, no history of pain, anorexia
Previous diseases	history of diabetes since 5 years, hypertension since 1 year back, history of accident faciomaxilery and left wrist joint dislocation 2 years back	No comorbidities History of an extract of a small lump 20 years back
Personal habits	No habits Sleep – adequate Appetite - normal Bowel, Bladder – Normal BP - 180/80 wt – 55 Kg SPO2 - 98	No Habits Sleep - disturbed Appetite - normal Bowel, Bladder - Normal BP - 110/70 wt – 75 Kg SPO2 - 96
Family history	Diabetes, hypertension	husband has carcinoma esophagus
Drug history	On medication for diabetes and hypertension	No drug history

An interpretability layer was added to improve the explainability of the diagnostic questions generated. Before a question is formed, the transformer's decoder extracts the attention distribution among patient input tokens. High-weight tokens, such as the length of the lump, family history, persistence of discomfort, and medication history, are used to visually highlight and display the significant clinical triggers. This offers a clear line of reasoning that can increase therapeutic acceptance and confidence by allowing physicians to identify the characteristics of the patient that contributed to each generated inquiry. For example, in a case where the patient reported noticing a developing lump without any discharge, the follow-up question generated by the model specifically addresses the patient's initial complaint that the lump was increasing but had not altered in size. Clinical data is easily understandable and less likely to be hesitantly used in the actual world because of its direct association with the context of the questions that were created.

IV. RESULTS AND DISCUSSIONS

The sample data, described in Table III, consists of 100 patients diagnosed with breast cancer, all married women, ensuring a uniform demographic base. Comorbidities were analyzed, finding that Hypertension (HTN) was present in 55% patients, and Diabetes Mellitus (DM) in 47% patients, making them the most common comorbidities. This was followed by Coronary Artery Disease (CAD) in 23% patients, Cerebrovascular Accident (CVA) in 14% patients, and Tuberculosis (TB) in 11% patients, indicating the significant presence of chronic diseases in the patient group. Lifestyle analysis showed that 96% had no addictions, while tobacco use was present in 3%, and alcohol use in 1% patients. The clinical presentation of breast cancer varied: lump presence was detected in 70%, pain was experienced by 30%, and discharge was reported by 3% patients.

These distributions highlight the need to take into account both clinical and lifestyle factors when designing the automated medical history collection model, allowing the generation of contextually relevant and patient-specific diagnostic questions. The dataset was partitioned into 80% training and 20% validation subsets to ensure balanced model evaluation. As summarized in Table III, the distribution of key comorbidities and lifestyle factors across the patient dataset highlights clinically relevant patterns that influence the design of the automated history-taking model.

TABLE III. CSET ATTRIBUTES ACROSS PATIENT RECORDS

Attributes	**Patient record attributes**		
	Category	**Count**	**Percentage (%)**
Sex	Female	100	100
Marital Status	Married	100	100
Comorbidities	HTN	55	55
	DM	47	47
	CAD	23	23
	TB	11	11
	CVA	14	14
Habits	Tobacco	3	3
	Alcohol	1	1
	None	96	96
Lump Presence	Lump	70	70
	No Lump	30	30
Pain Presence	Pain	30	30
	No Pain	70	70
Discharge	Discharge	3	3
	No Discharge	97	97

The performance of the proposed BreastCancerDiagNet framework was evaluated using Exploratory Data Analysis (EDA), convergence of model training, quantitative evaluation metrics, and qualitative evaluation of generated medical questions. The results demonstrate that the system is effective in producing context-aware, clinically relevant, and diagnostically meaningful questions, which directly contribute to automated medical history-taking.

Figure 2 shows the distributions of the pain-and-discharge symptoms in the records of the patient. These findings show that 30% of patients experienced pain, and most of them (70%) did not. Conversely, discharge was found in only 3% of all cases and 97% were found to have no discharge. These results indicate that pain is a medium-frequency symptom, and discharge is fairly uncommon, which highlights the variability of clinical presentation in the dataset.

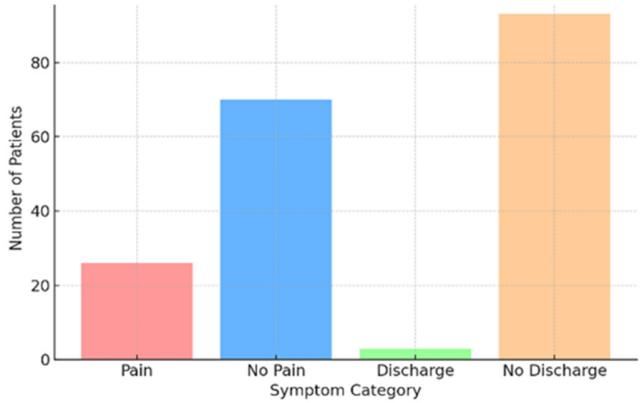

Fig. 2. Symptom counts for pain and discharge.

A. *Diagnostic Question Generation (QG) and Training Loss*

Diagnostic QG is the process by which the system creates meaningful and relevant questions for a patient by understanding both the written medical notes and structured clinical details, such as age, symptoms, and medical history. This helps doctors obtain a clearer picture of the patient's condition and ensures that important information is not missed. Training loss is a way to measure how far off the model's generated questions are from the correct ones during training. By keeping track of this loss, the model gradually learns to improve, producing more accurate and useful diagnostic questions over time.

Algorithm 1 describes how the BreastCancerDiagNet model creates diagnostic questions by looking at both patient text data and structured clinical information. First, it reads the patient's records using ClinicalBERT to understand important details in the text. At the same time, structured information, such as age, symptoms, and medical history, is processed separately. These two types of information are combined into a single context that the model uses to generate questions. During training, the model learns by comparing its questions with correct answers, and in practice, it can produce clear, patient-specific questions. Finally, the generated questions are organized into a structured questionnaire, and the model's performance is tracked to ensure that it works accurately.

```
Algorithm 1: QG process
# Step 1: Extract Text Features using
# ClinicalBERT
text_features =
  clinical_bert_encoder(input_ids,
  attention_mask)
# Step 2: Process Structured Data
struct_features =
  struct_mlp(structured_input)
# Step 3: Fuse Text and Structured
# Features
fused_features, context_vector =
  fuse(text_features, struct_features)
# Step 4: Generate Diagnostic Questions
logits = transformer_decoder(target_ids,
  context_vector)
# Step 5: Calculate Training Loss
loss = cross_entropy_loss(logits.view(-1,
  vocab_size), target_labels.view(-1))
# Step 6: Make Predictions in Inference
# Mode
generated_questions =
  generate_questions(fused_features)
# Step 7: Output Structured Diagnostic
# Questionnaire
final_questionnaire =
  tokenizer.decode(generated_questions,
  skip_special_tokens=True)
# Step 8: Monitor Performance Metrics
accuracy = compute_accuracy(predictions,
  ground_truth_labels)
```

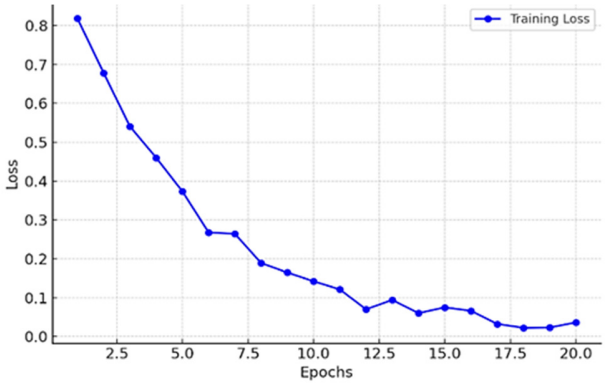

Fig. 3. Training loss curve showing convergence across epochs.

Figure 3 shows how the training loss changed over the course of 20 epochs during model training. In the beginning, the loss was quite high, around 0.82, but it decreased sharply during the initial stages of training. By the 8th epoch, the loss had already dropped significantly to approximately 0.20, indicating that the model was quickly learning from the data. After that point, the rate of decrease slowed down, and by the 15th epoch, the loss settled between 0.05 and 0.07. From there on, the curve remained stable without any signs of the loss increasing, showing that the model continued to learn effectively without overfitting the training data. This steady

reduction and eventual plateau of the loss suggest that the model successfully captured the important features from both structured and text-based patient data. As a result, the system is well-prepared to generate meaningful and relevant diagnostic questions, contributing to reliable automated medical history collection.

Figure 4 illustrates the results of the proposed BreastCancerDiagNet on the breast cancer dataset with three common measures of evaluation. The BLEU-4 score of the model was 0.42, which is a moderate overlap of the lexicon with the ground truth questions created by clinicians. The ROUGE-L score of 0.56 shows that the generated questions effectively extract the main clinical terms and phrases, whereas the BERT-F1 score of 0.88 indicates that the generated questions are semantically close to the reference questions, regardless of the difference in words. These findings indicate that the framework is capable of producing textually relevant questions that are semantically consistent with real clinical practice, making it robust in automated medical history-taking.

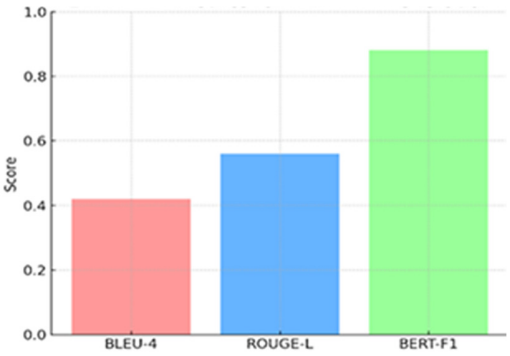

Fig. 4. Evaluation metrics for QG.

Table IV shows a comparison between the ground truth clinician-curated questions and those estimated by BreastCancerDiagNet, where it is clear that the framework repeatedly encapsulates the desired clinical meaning and that the phrasing is adjusted to patient-specific situations. Although the ground truth query on the lump presence examines whether the lump has altered in size, BreastCancerDiagNet extends the query to include the onset of the lump as well, and thus further details of the diagnosis. In the same vein, the pain-related question of the model focuses on duration, and the discharge-related question focuses on presence and duration, which match well with clinical reasoning. There are also questions about family history and drug use that display semantic fidelity, which are phrased by BreastCancerDiagNet in more approachable and context-sensitive modes.

These examples demonstrate that the developed questions retain the diagnostic purpose of the ground truth, as well as contribute to patient-centered communication, which supports the usefulness of the model in automated medical history-taking. The generated question patterns were cross-verified with standard breast cancer case-history templates reported in the literature, indicating initial clinical applicability pending formal clinician validation.

TABLE IV. EXAMPLES OF GROUND TRUTH VS. BREASTCANCERDIAGNET GENERATED QUESTIONS

QG comparison		
Clinical attribute	**Ground truth question**	**BreastCancerDiagNet generated question**
Lump presence	Has the lump changed in size over time?	When did you first notice the lump, and has it changed in size?
Pain	Does the patient have prolonged pain?	How long have you been experiencing pain in the breast?
Discharge	Is there any discharge from the breast?	Is there any nipple discharge, and what is its duration?
Family history	Family history of breast disease?	Do you have any family history of breast or related cancers?
Drug/ Reproductive	Drug history / OCP history?	Are you currently taking any medications or hormone therapy?
Lump presence	Has the lump changed in size over time?	When did you first notice the lump, and has it changed in size?

As shown in Table V, BreastCancerDiagNet consistently outperforms Seq2Seq and vanilla Transformer models across all evaluation metrics, achieving a BLEU-4 score of 0.42, a ROUGE-L of 0.56, and a BERT-F1 of 0.88. In addition, it demonstrates higher interpretability by generating context-aware and clinically meaningful questions, whereas baseline models tend to produce generic or incomplete outputs.

TABLE V. COMPARATIVE PERFORMANCE OF BREASTCANCERDIAGNET AGAINST BASELINE MODELS

Model	**BLEU-4**	**ROUGE-L**	**BERT-F1**	**Interpretability**
Seq2Seq (Attention)	0.31	0.44	0.75	Low (generic questions)
Vanilla Transformer	0.36	0.49	0.81	Moderate (semantic but incomplete)
BreastCancerDiagNet (Proposed)	0.42	0.56	0.88	High (clinically relevant)

V. CONCLUSION

BreastCancerDiagNet is a transformer-based framework for automated medical history-taking that can combine structured and unstructured patient information to create clinically relevant questions and summarize medical reports. With the help of ClinicalBERT embeddings, BiLSTM, self-attention, rotary embeddings, and FlashAttention, the framework was proven to be solid in quantitative measures and qualitative clinical analysis. Compared to baseline models, BreastCancerDiagNet generated more context-sensitive and diagnostic-relevant questions, and strong semantic alignment was assessed using BLEU, ROUGE, and BERTScore. The integration of RLHF provided flexibility and ongoing upgrades with clinician feedback as a much-needed step towards the balance between technical performance and clinical applicability. Integration with RLHF provided flexibility and ongoing upgrades with feedback provided by clinicians as a much-needed step towards the balance between technical performance and clinical applicability. These findings suggest that BreastCancerDiagNet can be used to reduce the duration of consultation, help early recognition of symptoms, and enhance overall diagnostic decision-making.

This study is restricted by a one-domain aspect of breast cancer and a small sample size. Future research will explore its expansion to multiple diseases and validation on large and heterogeneous clinical data to increase generalizability. Research is underway to expand the framework to multimodal clinical data, scale to larger disease subsets, and directly interface with EHR systems to deploy in real-time clinical settings.

REFERENCES

[1] K. Nassiri and M. A. Akhloufi, "Recent Advances in Large Language Models for Healthcare," *BioMedInformatics*, vol. 4, no. 2, pp. 1097–1143, Apr. 2024, https://doi.org/10.3390/biomedinformatics4020062.

[2] S. Chatterjee, A. Fruhling, K. Kotiadis, and D. Gartner, "Towards new frontiers of healthcare systems research using artificial intelligence and generative AI," *Health Systems*, vol. 13, no. 4, pp. 263–273, Oct. 2024, https://doi.org/10.1080/20476965.2024.2402128.

[3] H. Zhou *et al.*, "A Survey of Large Language Models in Medicine: Progress, Application, and Challenge." arXiv, July 23, 2024, https://doi.org/10.48550/arXiv.2311.05112.

[4] H. Yadav, P. Yadav, N. Yadav, and P. Chaudhary, "AI in Healthcare: A Survey on Medical Question Answering System," *South Eastern European Journal of Public Health*, pp. 1287–1298, Dec. 2024, https://doi.org/10.70135/seejph.vi.2683.

[5] G. Kell *et al.*, "Question answering systems for health professionals at the point of care—a systematic review," *Journal of the American Medical Informatics Association*, vol. 31, no. 4, pp. 1009–1024, Apr. 2024, https://doi.org/10.1093/jamia/ocae015.

[6] M. Sarrouti and S. O. El Alaoui, "SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions," *Artificial Intelligence in Medicine*, vol. 102, Jan. 2020, Art. no. 101767, https://doi.org/10.1016/j.artmed.2019.101767.

[7] H. Faris, M. Habib, M. Faris, A. Alomari, P. A. Castillo, and M. Alomari, "Classification of Arabic healthcare questions based on word embeddings learned from massive consultations: a deep learning approach," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 4, pp. 1811–1827, Apr. 2022, https://doi.org/10.1007/s12652-021-02948-w.

[8] S. Canchila, C. Meneses-Eraso, J. Casanoves-Boix, P. Cortés-Pellicer, and F. Castelló-Sirvent, "Natural language processing: An overview of models, transformers and applied practices," *Computer Science and Information Systems*, vol. 21, no. 3, pp. 1097–1145, 2024, https://doi.org/10.2298/CSIS230217031C.

[9] P. Rouzrokh *et al.*, "A Current Review of Generative AI in Medicine: Core Concepts, Applications, and Current Limitations," *Current Reviews in Musculoskeletal Medicine*, vol. 18, no. 7, pp. 246–266, Apr. 2025, https://doi.org/10.1007/s12178-025-09961-y.

[10] K. Singhal *et al.*, "Toward expert-level medical question answering with large language models," *Nature Medicine*, vol. 31, no. 3, pp. 943–950, Mar. 2025, https://doi.org/10.1038/s41591-024-03423-7.

[11] M. Cascella, F. Semeraro, J. Montomoli, V. Bellini, O. Piazza, and E. Bignami, "The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives," *Journal of Medical Systems*, vol. 48, no. 1, Feb. 2024, Art. no. 22, https://doi.org/10.1007/s10916-024-02045-3.

[12] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, https://doi.org/10.1093/bioinformatics/btz682.

[13] A. Chaddad, J. Peng, J. Xu, and A. Bouridane, "Survey of Explainable AI Techniques in Healthcare," *Sensors*, vol. 23, no. 2, Jan. 2023, Art. no. 634, https://doi.org/10.3390/s23020634.

[14] D. Wang and S. Zhang, "Large hlanguage models in medical and healthcare fields: applications, advances, and challenges," *Artificial Intelligence Review*, vol. 57, no. 11, Sept. 2024, Art. no. 299, https://doi.org/10.1007/s10462-024-10921-0.

[15] A. Nentidis, A. Krithara, K. Bougiatiotis, G. Paliouras, and I. Kakadiaris, "Results of the sixth edition of the BioASQ Challenge," in *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, Brussels, Belgium, 2018, pp. 1–10, https://doi.org/10.18653/v1/W18-5301.

[16] C. Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain, Apr. 2004, pp. 74–81.

[17] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Philadelphia, PA, USA, 2001, Art. no. 311, https://doi.org/10.3115/1073083.1073135.

[18] M. Reid, M. French, S. Andreopoulos, C. Wong, and N. Kee, "AI-generated multiple-choice questions in health science education: Stakeholder perspectives and implementation considerations," *Current Research in Physiology*, vol. 8, Jan. 2025, Art. no. 100160, https://doi.org/10.1016/j.crphys.2025.100160.

[19] S. Shen *et al.*, "On the Generation of Medical Question-Answer Pairs," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 8822–8829, Apr. 2020, https://doi.org/10.1609/aaai.v34i05.6410.

[20] Y. Cheng et al., "Guiding the Growth: Difficulty-Controllable Question Generation through Step-by-Step Rewriting," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 2021, pp. 5968–5978, https://doi.org/10.18653/v1/2021.acl-long.465.

[21] T. Searle, Z. Ibrahim, J. Teo, and R. J. B. Dobson, "Discharge summary hospital course summarisation of in patient Electronic Health Record text with clinical concept guided deep pre-trained Transformer models," *Journal of Biomedical Informatics*, vol. 141, May 2023, Art. no. 104358, https://doi.org/10.1016/j.jbi.2023.104358.

[22] P. Chakraborty, T. Chandrapragasam, A. Arunachalam, and S. Rafiammal, "Artificial Intelligence-based Oral Cancer Screening System using Smartphones," *Engineering, Technology & Applied Science Research*, vol. 13, no. 6, pp. 12054–12057, Dec. 2023, https://doi.org/10.48084/etasr.6364.

[23] T. Siddiqui, M. Latif, M. U. Farooq, M. A. Baig, and Y. S. Hassan, "Chronic Obstructive Pulmonary Disease Diagnosis with Bagging Ensemble Learning and ANN Classifiers," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14741–14746, June 2024, https://doi.org/10.48084/etasr.7106.