Curating Dataset Pipelines to Train Medical Chatbots on Early Sepsis Detection

Arup Das, Miriam Abecasis, Thomas Farrell, Isaac Sasson, Sophia Ramirez, Brooke Tortorelli, and Jiacun Wang

***Abstract*—Sepsis is a critical medical condition that arises when the body's response to infection causes life-threatening organ dysfunction. Despite increasing awareness and the use of protocol-driven management strategies, early diagnosis remains a persistent challenge in clinical practice, especially in high-pressure settings such as emergency departments and ICUs. Nurses, as first responders, are crucial in identifying early signs, but often work under cognitive overload and ambiguity of the protocol. Large language models (LLMs) represent frontier neural network techniques that use self-supervised learning algorithms to process and understand human languages or text. This work focuses on building a robust data gathering pipeline in order to ultimately create an interactive clinical chatbot fine-tuned on sepsis-specific knowledge. The pipeline consists of a three-step process, namely lexical analysis, semantic analysis, and Q&A quality evaluation, that utilizes artificial intelligence to collect training data in novel ways. It provides a feasible and cutting-edge framework for LLM-based chatbot design and development.**

***Key Words*—Sepsis, large language models, medical chatbot, lexical analysis, semantical analysis, LLM as a judge.**

I. Introduction

SEPSIS is a life-threatening condition that requires early detection and intervention to improve patient outcomes. Despite advances in medical protocols, early diagnosis remains challenging, particularly in high-stress environments such as emergency departments and ICUs. Nurses, who are often the first to interact with patients, play a crucial role in identifying early signs of sepsis but face significant cognitive load and ambiguity in protocols.

With the rapid development of artificial intelligence and data science [1, 2], especially Large Language Models (LLMs), new opportunities have emerged for the modernization and transmission of the diagnosis of medical problems [3]. With its powerful semantic understanding and knowledge reasoning capabilities, an LLM model could theoretically reduce the difficulty of understanding sepsis conditions, improve diagnostic decision-making efficiency, and thus effectively alleviate the challenges posed by long training cycles and the insufficient number of trained professionals. However, several major barriers block the effective application of LLMs in the sepsis domain.

These barriers include the scarcity of high-quality, curated sepsis-specific question–answer datasets, the risk of hallucinated or unsafe medical outputs, challenges in ensuring clinical safety and guideline adherence, high computational costs associated with large-scale models, and the cognitive complexity of sepsis presentations that demand precise, context-aware reasoning. Without addressing these limitations at the data and evaluation level, LLM-based systems risk producing unreliable or clinically inappropriate responses.

This study aims to address these challenges by providing a reliable, AI-driven tool that can assist healthcare professionals in real-time decision making. Using compact, yet powerful transformer models like Gemma 2B-IT, our chatbot aims to deliver accurate, clinically relevant responses to questions related to sepsis.

Artificial intelligence (AI) offers a promising avenue to support clinical decision-making in this context [4, 5, 6]. Large language models (LLMs), in particular, have demonstrated strong potential for synthesizing medical knowledge, answering clinician queries, and assisting with information retrieval. However, most existing models, such as BioGPT or Med-PaLM, are extremely resource-intensive, limiting their practicality for deployment in settings where lightweight, responsive systems are needed. Moreover, many prior efforts focus on demonstrating model performance while giving less attention to the foundational step of curating reliable datasets [7]. Without rigorous data pipelines, even the most advanced models risk producing clinically irrelevant or unsafe outputs.

Our work addresses this gap by focusing on the development of a robust data gathering and preprocessing pipeline tailored specifically for sepsis-related knowledge. Rather than beginning with model fine-tuning, we prioritize the careful curation of a structured dataset sourced from medical literature. By implementing a multi-step pipeline consisting of lexical

Manuscript received December 3, 2025; revised December 11 and December 18, 2025; accepted January 6, 2025. This article was recommended for publication by Associate Editor Shujin Qin upon evaluation of the reviewers' comments.

This work was supported by the School of Science Summer Research Fund at Monmouth University.

A. Das is with the Department of Computer Science and Software Engineering, West Long Branch, NJ 07764, USA (e-mail: adas@monmouth.edu).

M. Abecasis is with the Department of Computer Science and Software Engineering, West Long Branch, NJ 07764, USA (e-mail: s1354404@monmouth.edu).

T. Farrell is with the Department of Computer Science and Software Engineering, West Long Branch, NJ 07764, USA (e-mail: s1340394@monmouth.edu).

I. Sasson is with the Department of Computer Science and Software Engineering, West Long Branch, NJ 07764, USA (e-mail: s1374593@monmouth.edu).

S. Ramirez is with the Department of Computer Science and Software Engineering, West Long Branch, NJ 07764, USA (e-mail: s1347596@monmouth.edu).

B. Tortorelli is with the Department of Computer Science and Software Engineering, West Long Branch, NJ 07764, USA (e-mail: s1305131@monmouth.edu).

J. Wang is with the Department of Computer Science and Software Engineering, West Long Branch, NJ 07764, USA (e-mail: jwang@monmouth.edu).

Corresponding author: Sophia Ramirez

analysis, semantic analysis, and LLM-as-a-Judge, we establish a reproducible workflow that emphasizes data quality, safety, and clinical relevance.

The main contributions of the work are as follows.

(1) A three-step data curation pipeline is developed that utilizes advanced large language models to compile a data set to train a chatbot.

(2) Lexical Analysis: The first step in our pipeline during which question-answer pairs are analyzed based on lexical similarity, and pairs that are deemed too similar to others in the set are discarded.

(3) Semantic Analysis: In this step, the question-answer pairs that remain after being passed through the lexical analysis filter are analyzed on the basis of semantic similarity. Once again, we remove pairs deemed similar to others in the set.

(4) Q&A Quality Assessment: In the last step in our pipeline, we use large language models to evaluate the remaining question-answer pairs (LLM as a judge) based on metrics such as quality, accuracy, and clinical relevance.

II. Related Work

Previous research has explored the use of large language models in healthcare, such as BioGPT [8] and Med-PaLM [9]. However, these models often require substantial computational resources, limiting their deployment in resource-constrained settings. Recent advancements in LoRA-based fine-tuning (e.g., QLoRA [10]) and LLM-as-a-Judge methodologies (e.g., KoalaEval) have paved the way for more efficient, clinically viable workflows. Our study builds on these innovations to create a specialized, deployable solution for sepsis management using lightweight models. In contrast to these large, resource-intensive systems, we present a lightweight and reproducible dataset curation pipeline that runs on accessible hardware and filters redundancy and safety risks before fine-tuning.

A. Sepsis Prediction

Early and precise recognition of sepsis is critical as delayed treatment increases the mortality rate dramatically. Srimedha et al. proposed a classifier in [11] that can accurately predict sepsis up to six hours before the disease is clinically diagnosed. The predictor utilizes a patient's electronic medical records, demographics, and vital signs. In [12], Lyra et al. optimized and evaluated four prediction models with different architectural concepts. Two public datasets containing clinical data from adults and neonates were used for training. In [13], Dai et al. uses a deep reinforcement learning framework for solving early prediction of sepsis. To improve the prediction performance of sepsis, Apalak et al. [14] used conditional recurrent adversarial networks, which is trained with the output of a conditional GAN and evaluated on an unseen dataset. The same authors present an early sepsis detection algorithm utilizing the Medical Information Mart for Intensive Care (MIMIC-III) Clinical Dataset and MIMIC-Waveform Database. The algorithm utilizes ECG signals, as part of a patient monitoring system for individuals in ICU [15]. In [16], a multilayer machine learning approach is presented to analyze continuous high-frequency data, which has the ability to detect early patients at risk of sepsis. Most recently, Giordano et al. introduces SepAl, an energy-efficient and lightweight neural network, using only data from low-power wearable sensors and body temperature sensors to deliver sepsis alerts in real time [17].

B. LLMs in Healthcare

In [3], AcuGPT-agent, a novel intelligent system powered by a domain-specific large language model (LLM) designed for acupuncture-based infertility treatment is presented. It works as a chatbot. A similar work is reported by Griot et al [18], which presents a secure, fully onpremises, GDPR-compliant LLM chatbot integrated into the Epic EHR system at a European university hospital. Wu et al. introduce a novel LLM- and applied it to the diagnosis of brain tumors in healthcare informatics [19]. In [20], a multimodal medical chatbot that leverages Gemini-2.0-Flash Model alongside a novel RAG architecture to support preliminary medical diagnosis and recommendations is presented. The system integrates textual prompt analysis and medical image interpretation. Liu et all [21] provides a survey on current psychiatric practice of LLMs, along with a series of corpus resources that could be used for training psychiatric LLMs. Limitations concerning LLM reproducibility, capabilities, usability, interpretability in clinical settings are discussed, along with ethical concerns.

C. Q&A Pairs for Chatbot

A chatbot system that offers medical consultation services to patients with ophthalmologic diseases is presented in [22], in which the QA dataset is created closely with an ophthalmologist to obtain and verify medical data. In [23], Calfoforo et al. investigates the integration of Retrieval-Augmented Generation (RAG) and the LangChain framework to develop a QA system using the Llama-2 large-language model. The QA system was designed to improve information retrieval accuracy and relevance for policy-related questions of a university. The QLoRA technique was employed for model fine-tuning. Rasool et al. assess LLM performance in QA types, including single choice, yes-no, multiple choice, and number extraction questions from documents [24]. An automated disease QA system named Disease Guru–Long-Form Question Answer (DG-LFQA) is introduced by Sukhwal et. al. in [25]. DG-LFQA employs LLM and knowledge graph to answer disease-related questions appropriate for users.

III. Dataset Curation

The quality of training datasets is critical to any machine learning mission [26, 27]. We curated a high-quality dataset by filtering relevant questions and answers about sepsis from PubMedQA and HealthCareMagic QA. The process involved:

1. Article Search and Curation: We searched for relevant articles on sepsis and compiled them into curated "golden PDFs," excluding irrelevant information.

2. PDF Processing: We used the Gemma-2B model to process these PDFs, converting them to text and chunking them into manageable sections.

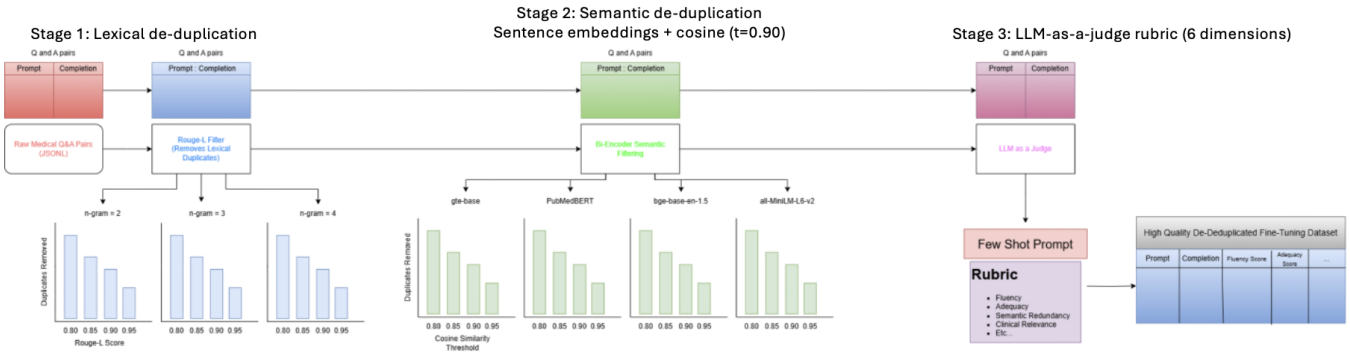


Fig. 1. The data pipeline. It summarizes the full curation workflow: (i) lexical deduplication using ROUGE-L with n-grams, (ii) semantic deduplication using embedding similarity with cosine thresholding, and (iii) final quality screening using an LLM-as-a-judge rubric focused on accuracy, helpfulness, and safety.

3. Q&A Generation: Each chunk was processed to generate Q&A pairs, formatted into JSON lines for easy integration into our dataset.

4. Data Compilation: The generated pairs were collected into a DataFrame and saved as a JSONL file, ensuring a robust and well-structured dataset.

To optimize our three-stage pipeline, we conducted systematic experiments at each filtering stage. For lexical analysis, we tested multiple n-gram configurations (n=2, 3, and 4) combined with four ROUGE-L similarity thresholds (0.80, 0.85, 0.90, and 0.95) to identify the optimal balance between duplicate removal and content preservation. In the semantic analysis stage, we evaluated four different embedding models—`PubMedBERT-SBERT`, `all-MiniLM-L6-v2`, `gte-base`, and `bge-base-en-v1.5`—across three cosine similarity thresholds (0.85, 0.90, and 0.95) to determine which combination best captured medical semantic similarity. For the LLM-as-a-Judge component, we employed `MedGemma 4B-IT` with a six-dimensional evaluation rubric, testing different quantization and generation parameters to balance evaluation quality with computational efficiency. The following subsections detail the results and final parameter selections for each stage.

As illustrated in Figure 1, the pipeline consists of lexical, semantic, and LLM-based filtering stages.

A. *Lexical Analysis*

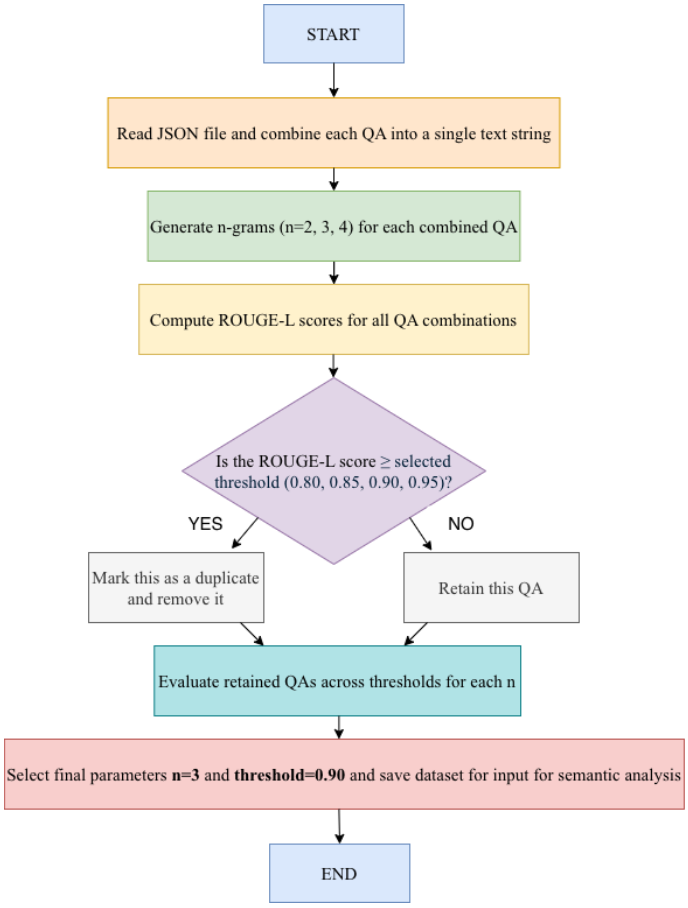


Fig. 2. Lexical analysis pipeline.

As shown in the pipeline in Figure 2, we performed a lexical analysis of the QA pairs to prepare for the semantic analysis. Lexical analysis is the stage of text processing in which characters are divided into tokens, such as words or symbols, while leaving out irrelevant elements such as spaces [28]. The input for the beginning of the analysis was the combined pairs of questions and answers in one text string to compare pairs based on the exact use of words. Duplicates at the surface level

were removed through lexical filtering and approximately 3.2 million pairwise comparisons were evaluated from 2554 QAs.

A combination of ROUGE-L threshold tuning and n-grams was utilized to capture phrasing and structural repetition beyond simple word overlap. ROUGE-L measured text similarity based on word order using the longest common subsequence (LCS), with scores ranging from 0 to 1 [29]. ROUGE-L scores preserve sentence structure by only considering ordered word overlaps [30]. As defined by Lin, given two sequences (X) and (Y) of lengths (m) and (n), recall and precision are defined as

$$R_{lcs} = \frac{LCS(X,Y)}{m} \tag{1}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \tag{2}$$

and combined into an F-measure:

$$F_{lcs} = \frac{(1+\beta^2)\, R_{lcs}\, P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}, \tag{3}$$

which is the similarity calculation [30]. In our code, we utilized the rouge-scorer implementation from Google's rouge-score package. In this case, β is fixed at 1 [31].

As previously stated, ROUGE-L considers the longest common subsequence, which means that it could undervalue shorter, similar fragments. Therefore, to establish a thorough lexical filter, we incorporated n-grams to capture smaller subsequence overlaps. N-grams break each question into short phrases (n = 2, 3, and 4) to identify repeated phrasing [32]. ROUGE-L thresholds of 0.80, 0.85, 0.90, and 0.95 were tested for each n-gram setting to determine whether two QAs were similar enough for one to be discarded. After testing all combinations of n and the threshold in the aforementioned range, we noticed that there was no substantial difference in the number of questions-answer pairs retained. Therefore, we considered standard practices when making our decisions pertaining to the ROUGE scores. Trigram overlap considers greater context than bigram overlap. However, higher-order n-grams, such as n = 4 tend to fail at detecting near-duplicate words if there are phrasing discrepancies [33]. Thus, to establish balance, we decided to fix n = 3.

Given n = 3, the threshold needed to be set accordingly for preprocessing. Choosing 0.8 or 0.85 is too limiting in this phase and removes too many phrases with meaningful differences. However, a threshold of 0.95 retained several near-identical pairs. Thus, n = 3 and a threshold of = 0.9 established a balance for this initial phase. In this case, we were able to remove explicit redundancy while retaining meaningful differences.

This approach is derived from previous work that established that balance prevents semantic loss, which is especially important given that the tuned dataset of 2316 question–answer pairs was used as input for subsequent semantic analysis [34].

B. Semantic Analysis

Following lexical filtering, we conducted a semantic analysis to remove question–answer pairs that exhibited excessive similarity in content, as opposed to just lexical similarity. Before now, we were only removing question-answer pairs when the wording being used was too similar. Now, however, we are looking at the actual meaning of words and sentences in context and removing question-answer pairs based on contextual similarity.

Each concatenated pair (qa_text = "Q: <question>A : <answer>") was then embedded using PubMedBERT-SBERT (NeuML/pubmedbert-base-embeddings) via the Sentence-Transformers framework. Sentence-BERT (SBERT) processes each input by first segmenting the text into subword tokens, allowing rare or unfamiliar words to be effectively represented. These tokens are then passed through a BERT-based transformer, where self-attention constructs context-aware embeddings for each token. To derive a fixed-length representation of the entire question-answer pair, SBERT applies a pooling operation across the token embeddings, resulting in a 768-dimensional sentence vector.

We performed encoding in mini-batches with the setting `normalize_embeddings = True`, ensuring that each vector lay on the unit sphere so that cosine similarity was equivalent to the inner product. For comparison baselines, we applied the same pipeline using three widely adopted general-purpose encoders: `all-MiniLM-L6-v2`, `gte-base`, and `bge-base-en-v1.5`. A FAISS `IndexFlatIP` was then constructed over the 768-dimensional `float32` embeddings, and for each item i, we retrieved all neighbors. A pair (i, j) (excluding self-matches) was considered a near-duplicate if its cosine similarity exceeded the threshold τ. To ensure determinism in a single pass, only the higher index item $(j > i)$ was removed when a duplicate was identified. This procedure yields an order-stable greedy deduplication strategy: retain the first occurrence and eliminate subsequent entries that are semantically too close. The comparative results for each encoder are shown in Figures 2-5.

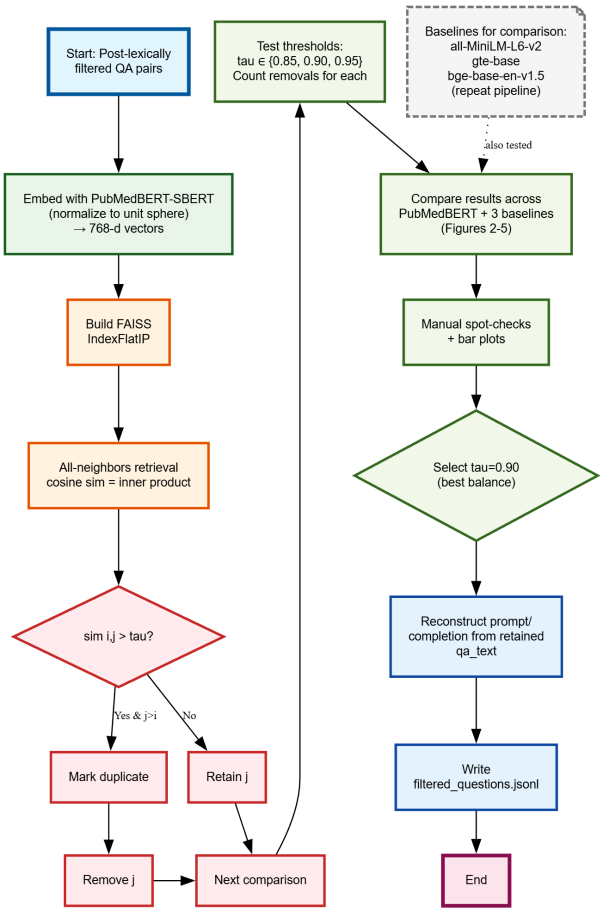


Fig. 3. Semantic Pipeline.

This representation enables comparison of question-answer pairs based on semantic meaning rather than exact wording. As a result, two pairs that differ noticeably in phrasing can still be identified as equivalent–or sufficiently similar to warrant removal–if their underlying meaning overlaps too closely [35].

We evaluated thresholds $\tau \in \{0.85, 0.90, 0.95\}$ by counting the number of pairs that would be removed at each value and visualizing the results with a bar plot (Figures 2-5). In this procedure, every question-answer pair was compared against all others, and each pairwise comparison was assigned a similarity score between 0 and 1, where 0 indicated no similarity and 1 indicated an exact match. Our objective was to eliminate pairs whose semantic similarity exceeded the threshold, and thus we tested multiple τ values to assess the trade-off. For example, a pair with similarity score of 0.89 would be removed under $\tau = 0.85$ but retained under $\tau = 0.90$.

Based on these results and manual spot-checks (Figures 2-5), we found that $\tau = 0.90$ offered the best balance. A threshold of 0.85 proved too aggressive, as it removed paraphrases that contributed meaningful clinical nuance, whereas 0.95 was too permissive, allowing clear rephrasings to remain. We therefore adopted $\tau = 0.90$, after which we reconstructed the original prompt and completion fields from the retained `qa_text` and saved the filtered dataset as `filtered_questions.jsonl`.

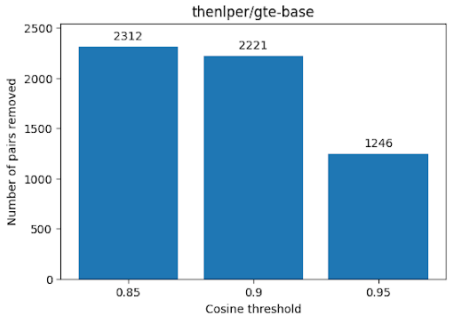


Fig. 4. Evaluation of general-purpose encoder thenlper/gte-base at different cosine thresholds.

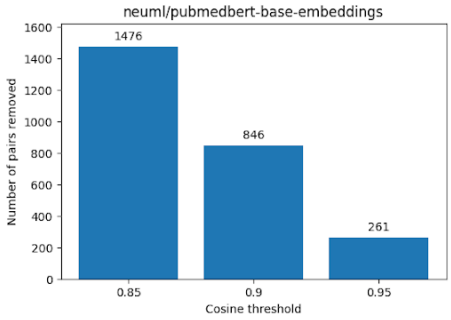


Fig. 5. Evaluation of general-purpose encoder neuml/pubmedbert-base-embeddings at different cosine thresholds.

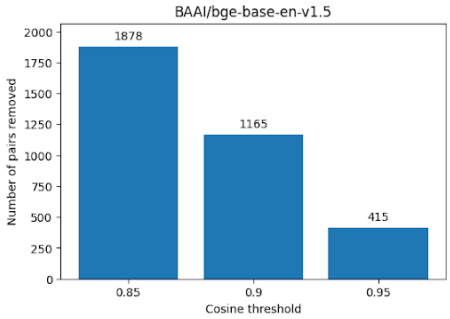


Fig. 6. Evaluation of general-purpose encoder BAAI/bge-base-en-v1.5 at different cosine thresholds.

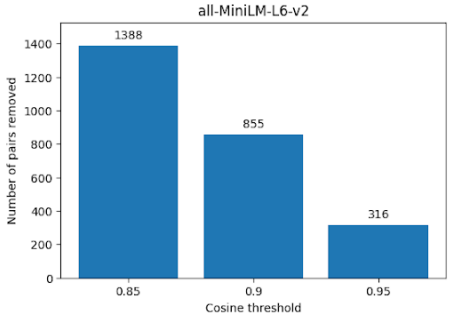


Fig. 7. Evaluation of general-purpose encoder all-miniLM-L6-v2 at different cosine thresholds.

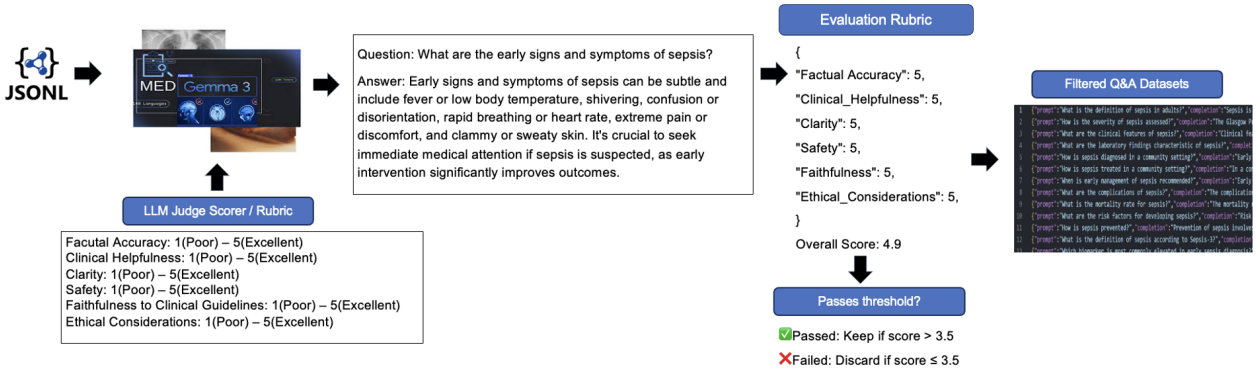


Fig. 8. LLM as a Judge.

C. LLM as a Judge

The concept of LLM as a Judge refers to the use of advanced large language models to automatically evaluate the quality, accuracy, and clinical relevance of text-based responses—such as question–answer (Q&A) pairs—in a structured, objective, and scalable manner. Traditionally, such evaluation has depended on domain experts, typically clinicians or medical researchers, who manually review and score each response. While expert judgment provides high credibility, it introduces several well-known limitations. Manual review is time-consuming, as each Q&A pair must be individually read, analyzed, and rated, often requiring hours or even days to process large datasets. The process is also subjective: inter-rater reliability tends to be low for nuanced or ambiguous cases, since judgments may vary depending on a reviewer's clinical experience, interpretation, or personal biases. Furthermore, manual annotation is not scalable. Evaluating thousands of pairs becomes prohibitively expensive and slow, rendering real-time or large-scale deployment impractical. The process is also resource-intensive, requiring ongoing recruitment, training, and management of annotators, which places financial and logistical strain on projects. Finally, the availability of expert reviewers is often limited, leading to bottlenecks or delays, particularly in time-sensitive scenarios such as clinical trials or urgent decision-support systems.

In contrast, LLM-based evaluation addresses these constraints by introducing speed, consistency, and scalability. Once deployed, an LLM can process thousands of Q&A pairs in minutes, enabling rapid feedback loops and real-time integration into research pipelines. Evaluations are applied uniformly according to predefined criteria, which reduces subjectivity and ensures reproducibility across different datasets and projects. The method is also inherently scalable, capable of handling datasets ranging from hundreds to millions of examples with relatively modest computational costs. Importantly, LLM-based evaluation is resource-light and always available: it does not require continual recruitment of experts, it operates continuously without fatigue, and it can provide on-demand evaluations around the clock.

For this study, we employ MedGemma 4B-IT, a lightweight yet clinically optimized large language model fine-tuned for medical applications. Its domain specialization makes it particularly well-suited for assessing sepsis-related Q&A pairs, ensuring that evaluations capture not only general linguistic quality but also clinical accuracy, safety, and alignment with medical guidelines.

To ensure comprehensive and fair evaluation, we designed a multi-dimensional rubric that assesses Q&A pairs across six dimensions: factual accuracy, clinical helpfulness, clarity, safety, faithfulness to clinical guidelines, and ethical considerations. Each criterion is scored on a five-point Likert scale, with 1 representing "poor" and 5 representing "excellent." Factual accuracy evaluates whether an answer is medically correct and free from errors, such as providing accurate descriptions of the signs and treatment of sepsis. Clinical helpfulness measures whether the answer provides clear, relevant, and contextually appropriate information that supports understanding without offering diagnostic or therapeutic advice. Clarity assesses the ease with which the information can be understood. Safety checks for potentially harmful or misleading recommendations, for instance suggesting unsupervised antibiotic use. Faithfulness to guidelines ensures that responses remain consistent with established standards of care, such as WHO or NICE recommendations. Finally, ethical considerations evaluate fairness, non-discrimination, and respect for patient autonomy and confidentiality.

a) Safety Definition and Failure Modes.: In this work, a response is considered *unsafe* if it exhibits any of the following behaviors: (1) hallucinated clinical information, including fabricated symptoms, laboratory values, diagnoses, or treatments not supported by the source document; (2) provision of diagnostic or therapeutic advice, such as recommending medications, interventions, or definitive diagnoses, rather than informational guidance; (3) omission of uncertainty or escalation cues in high-risk scenarios where consultation with a qualified healthcare professional is warranted; (4) contradiction of the source document or established clinical knowledge; or (5) false reassurance in the presence of potentially life-threatening conditions such as sepsis. Only Q&A pairs that avoid these behaviors and adhere to conservative, non-actionable clinical language are retained.

During evaluation, the LLM-as-a-judge explicitly screens for common failure modes observed in medical language model outputs, including hallucinated or unverifiable clinical

claims, overconfident diagnostic statements, partial responses that omit critical safety warnings, inconsistent reasoning across similar clinical contexts, and fluent but factually incorrect explanations.

For each candidate Q&A pair, MedGemma 4B-IT evaluates (i) factual alignment with the source document, (ii) clinical appropriateness of scope, and (iii) compliance with the safety criteria defined above. Q&A pairs failing any of these criteria are discarded, reflecting a conservative filtering strategy that prioritizes risk minimization over dataset size.

To anchor model evaluations to this rubric, we employ few-shot prompting, which supplies the model with annotated examples of high- and low-quality Q&A pairs. High-scoring examples demonstrate comprehensive, accurate, and clinically safe responses.

```
    few_shot_examples = """<start_of_turn>user
Example 1 (High Score)
Question: What are the early signs and symptoms of sepsis?
Answer: Early signs and symptoms of sepsis can be subtle and include fever or
low body temperature, shivering, confusion or disorientation, rapid breathing
or heart rate, extreme pain or discomfort, and clammy or sweaty skin.
It's crucial to seek immediate medical attention if sepsis is suspected,
as early intervention significantly improves outcomes.
<end_of_turn>
<start_of_turn>assistant
Evaluation:
{
  "Factual_Accuracy": 5,
  "Clinical_Helpfulness": 5,
  "Clarity": 5,
  "Safety": 5,
  "Faithfulness": 5,
  "Ethical_Considerations": 5,
}
Overall Score: 4.9
<end_of_turn>
```

Fig. 9. Example of high score.

```
<start_of_turn>user
Example 2 (Low Score)
Question: How is sepsis treated?
Answer: You can just take some antibiotics you have at home if you feel sick
with a fever. Drink lots of fluids.
<end_of_turn>
<start_of_turn>assistant
Evaluation:
{
  "Factual_Accuracy": 1,
  "Clinical_Helpfulness": 1,
  "Clarity": 3,
  "Safety": 1,
  "Faithfulness": 1,
  "Ethical_Considerations": 1,
}
Overall Score: 1.6
<end_of_turn>"""
```

Fig. 10. Example of low score.

For instance, when asked "What are the early signs and symptoms of sepsis?", a high-quality response correctly identified early indicators such as fever, shivering, confusion, rapid breathing, and clammy skin, while emphasizing the urgency of seeking immediate medical attention. This example received high scores across all dimensions, reflecting its factual accuracy, clarity, and clinical helpfulness. In contrast, low-scoring examples highlight typical failure modes. When asked "How is sepsis treated?", a poor-quality response suggested taking leftover antibiotics at home and drinking fluids. Such advice is unsafe, misleading, and inconsistent with clinical best practices, resulting in low scores across factual accuracy, safety, and faithfulness. These examples not only demonstrate the application of the rubric but also provide anchors that help the model generalize consistently to unseen Q&A pairs.

The evaluation pipeline was implemented with MedGemma 4B-IT configured in four-bit quantization (NF4, BFloat16) to reduce memory usage while preserving evaluation fidelity. All evaluations were conducted on a Google Colab T4 GPU with 16 GB of VRAM, operating in evaluation mode to ensure stability. Inputs were tokenized to a maximum sequence length of 8192 tokens, with padding and truncation enabled as needed. For generation, we set `max_new_tokens` to 200, with a sampling temperature of 0.7 and a top-k value of 40, balancing diversity with determinism in outputs. Evaluations were batched in groups of eight using a dedicated `QAPairEvaluator` class, which parsed outputs into structured JSON evaluations aligned with the six rubric dimensions.

This configuration proved both efficient and practical. Each batch of eight Q&A pairs required approximately 12 seconds to process, yielding an effective throughput of roughly 96 pairs per minute. Peak GPU memory usage was 10.2 GB under four-bit quantization, a substantial reduction compared to the 32 GB required for full-precision inference. These results demonstrate that reliable, rubric-based evaluation of medical Q&A datasets can be achieved on accessible hardware, making LLM-based judgment both scalable and cost-effective.

b) Limitations.: We acknowledge that reliance on a single LLM-based judge introduces limitations, including potential model bias and incomplete coverage of rare or edge-case failure modes. While prior studies have demonstrated the feasibility of LLM-based evaluators in domain-specific and medical settings, future work will explore multi-judge ensembles, cross-model agreement, and clinician-in-the-loop validation to further strengthen safety assurances.

IV. Conclusion

The rapid development of artificial intelligence and data science has provided a new avenue for disease diagnosis. Our work introduces a practical, open-source pipeline for creating compact, clinically relevant chatbots. The methodology can extend to other medical verticals such as cardiology, dermatology, or patient education, demonstrating the versatility and potential impact of lightweight LLMs in healthcare. Future work will include the application of explainable AI in the design and development of medical chatbots [36, 37], which will increase the trustworthiness of the AI tool.

References

[1] K. Zhang, Y. Xiao, J. Wang, M. Du, X. Guo, R. Zhou, C. Shi, and Z. Zhao, "DP-GAN: A transmission line bolt defects generation network based on dual discriminator architecture and pseudo-enhancement strategy," *IEEE Transactions on Power Delivery*, vol. 39, no. 3, pp. 1622–1633, 2024.

[2] X. Guo, C. Jiao, J. Wang, S. Qin, B. Hu, L. Qi, X. Lang, and Z. Zhang, "LLM-assisted reinforcement learning for u-shaped and circular hybrid disassembly line balancing in iot-enabled smart manufacturing," *Electronics*, vol. 14, no. 11, 2025.

[3] J. Wen, D. Liu, Y. Xie, Y. Ren, J. Wang, Y. Xia, and P. Zhu, "AcuGPT-agent: An LLM-powered intelligent system for acupuncture-based infertility treatment," *Neurocomputing*, vol. 652, p. 131116, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231225017886

[4] W. Qu, L. Zheng, D. Wang, J. Wang, and H. Pan, "Time-aware transformer-based prediction model for aecopd," *Studies in Health Technology and Informatics*, Aug. 2025.
[5] Y. Hu, L. Zheng, and J. Wang, "Predicting ICU length of stay for patients with diabetes using machine learning techniques," in *2022 International Conference on Cyber-Physical Social Intelligence (ICCSI)*, 2022, pp. 417–422.
[6] C. Xu, P. Zhu, J. Wang, and G. Fortino, "Improving the local diagnostic explanations of diabetes mellitus with the ensemble of label noise filters," *Information Fusion*, vol. 117, p. 102928, 2025.
[7] J. Zhou, J. Wang, and J. Wang, "A simulation engine for stochastic timed petri nets and application to emergency healthcare systems," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 4, pp. 969–980, 2019.
[8] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," November 2022.
[9] K. Singhal, S. Azizi, T. Tu *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.
[10] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLORA: efficient finetuning of quantized llms," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.
[11] B. C. Srimedha, R. Naveen Raj, and V. Mayya, "A comprehensive machine learning based pipeline for an accurate early prediction of sepsis in icu," *IEEE Access*, vol. 10, pp. 105 120–105 132, 2022.
[12] S. Lyra, J. Jin, S. Leonhardt, and M. Lüken, "Early prediction of neonatal sepsis from synthetic clinical data using machine learning," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2023, pp. 1–4.
[13] H. Dai, H.-G. Hwang, and V. S. Tseng, "Poems: Policy network-based early warning monitoring system for sepsis in intensive care units," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 7, pp. 3610–3621, 2023.
[14] M. Apalak and K. Kiasaleh, "Improving sepsis prediction performance using conditional recurrent adversarial networks," *IEEE Access*, vol. 10, pp. 134 466–134 476, 2022.
[15] ——, "Advancing early detection of sepsis with temporal convolutional networks using ecg signals," *IEEE Access*, vol. 12, pp. 3417–3427, 2024.
[16] F. van Wyk, A. Khojandi, and R. Kamaleswaran, "Improving prediction performance using hierarchical analysis of real-time data: A sepsis case study," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 978–986, 2019.
[17] M. Giordano, K. Dheman, and M. Magno, "SepAl: Sepsis alerts on low-power wearables with digital biomarkers and on-device tiny machine learning," *IEEE Sensors Journal*, vol. 25, no. 5, pp. 7858–7866, 2025.
[18] M. Griot, J. Vanderdonckt, and D. Yuksel, "Implementation of large language models in electronic health records," *Research Square Preprint*, July 2025.
[19] D. Wu, L. Nie, R. A. Mumtaz, and K. Agarwal, "A LLM-based hybrid-transformer diagnosis system in healthcare," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 9, pp. 6428–6439, 2025.
[20] I. Agarwal, V. Sakthivel, and P. Prakash, "Toward inclusive healthcare: An LLM-based multimodal chatbot for preliminary diagnosis," *IEEE Access*, vol. 13, pp. 136 420–136 432, 2025.
[21] Z. Liu, Y. Bao, S. Zeng, R. Qian, M. Deng, A. Gu, J. Li, W. Wang, W. Cai, W. Li, H. Wang, D. Xu, and G. N. Lin, "Large language models in psychiatry: Current applications, limitations, and future scope," *Big Data Mining and Analytics*, vol. 7, no. 4, pp. 1148–1168, 2024.
[22] J. H. Lee, M.-S. Jeong, J.-U. Cho, H.-K. Jeon, J.-H. Park, K.-D. Shin, S.-J. Song, and Y.-G. Cheong, "Developing a ophthalmic chatbot system," in *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 2021, pp. 1–7.
[23] J. C. Calfoforo and R. C. Raga, "Unleashing ai in education: A pre-trained llms for accurate and efficient question-answering systems," in *2024 21st International Conference on Information Technology Based Higher Education and Training (ITHET)*, 2024, pp. 1–6.
[24] Z. Rasool, S. Kurniawan, S. Balugo, S. Barnett, R. Vasa, C. Chesser, B. M. Hampstead, S. Belleville, K. Mouzakis, and A. Bahar-Fuchs, "Evaluating llms on document-based qa: Exact answer selection and numerical extraction using cogtale dataset," *Natural Language Processing Journal*, vol. 8, p. 100083, 2024.
[25] P. C. Sukhwal, V. Rajan, and A. Kankanhalli, "A joint LLM-KG system for disease Q&A," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 3, pp. 2257–2270, 2025.
[26] H. Yang, J. Wang, and J. Wang, "Efficient detection of forest fire smoke in uav aerial imagery based on an improved yolov5 model and transfer learning," *Remote Sensing*, vol. 15, no. 23, 2023.
[27] Y. Tian, G. Liu, J. Wang, and M. Zhou, "Asa-gnn: Adaptive sampling and aggregation-based graph neural network for transaction fraud detection," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 3, pp. 3536–3549, 2024.
[28] A. W. Appel, *Modern Compiler Implementation in Java*. Cambridge University Press, 1998.
[29] N. Bui, G. Nguyen, N. Nguyen, B. Vo, L. Vo, T. Huynh, A. Tang, V. N. Tran, T. Huynh, H. Q. Nguyen, and M. Dinh, "Fine-tuning large language models for improved health communication in low-resource languages," *Computer Methods and Programs in Biomedicine*, vol. 263, p. 108655, 2025.
[30] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proceedings of the ACL Workshop on Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.
[31] G. Research, "Rouge-score: A python implementation of the rouge metric," https://github.com/google-research/google-research/tree/master/rouge, 2019, accessed: 2025-10-03.
[32] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–394, 1999.
[33] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the ACL Workshop on Text Summarization Branches Out*, 2004, pp. 74–81.
[34] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 1073–1083.
[35] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese bert-networks," 2019. [Online]. Available: https://arxiv.org/abs/1908.10084
[36] H. Han, W. Li, J. Wang, G. Qin, and X. Qin, "Enhance explainability of manifold learning," *Neurocomputing*, vol. 500, pp. 877–895, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222007044
[37] H. Han, Y. Wu, J. Wang, and A. Han, "Interpretable machine learning assessment," *Neurocomputing*, vol. 561, p. 126891, 2023.

Arup Das is a leading expert in Artificial Intelligence and Machine Learning with extensive experience in applied AI, generative AI, and data-driven innovation across industries. He has held senior roles at UiPath, Avenue One, Compass, and Machine Analytics, where he led the development of AI-driven systems for sectors such as financial services, healthcare, and manufacturing. He holds advanced degrees from Cornell University, Villanova University, and Stony Brook University, and is co-author of The Generative AI Practitioner's Guide.

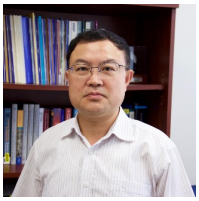
Sophia Ramirez is pursuing a Bachelor of Science in Software Engineering at Monmouth University, NJ, United States. Her research interests include machine learning, large language models, and their applications in healthcare. She completed an internship at TerraCycle, where she gained experience in information technology and project management. Sophia is passionate about leveraging artificial intelligence to improve early disease detection and patient outcomes, with a particular focus on developing ethical and interpretable AI tools.

Issac Sasson is pursuing a Bachelor of Science in Computer Science with minors in Mathematics and Data Science at Monmouth University, NJ, United States. His research interests include artificial intelligence, mobile app development, and game design. He completed an internship with the Monmouth Youth Communication Club, where he gained valuable experience in project management and full-stack development. He has won multiple hackathons and Capture the Flag (CTF) competitions.

Miriam Abecasis is pursuing a Bachelor of Science in Software Engineering and a Bachelor of Science in Mathematics at Monmouth University, NJ, United States. Her research interests include large language models, quantum graphs and their applications at a macroscopic level, virtual reality and how it could be incorporated in the classroom, and quantum computing. She has received several grants for her work in quantum graphs and recently received a Mathematical Association of America (MAA) Outstanding Poster Award.

Brooke Tortorelli is pursuing a Master of Science in Data Science at Monmouth University. She received her BS in Mathematics with a Concentration in Statistics and BA in Music with a Concentration in Musical Theatre from Monmouth University in 2024. During her undergraduate years, she proved that Brahmagupta quadrilaterals are lattice and published the result in *Involve, A Journal of Mathematics*. Additionally, she studied applied mathematics in healthcare, regarding SIR models and their parameters. Her research contributions awarded her the Monmouth University Faculty Recognition Award. Currently, she is applying natural language processing techniques to speech-language pathology.

Thomas Farrell is pursuing a Bachelor of Science in Computer Science with a minor in Mathematics at Monmouth University. His research interests include robotics, artificial intelligence, and advanced prompt-engineering methods for large language models. He is particularly interested in the development of evaluation frameworks for AI systems, including rubric design, weight-scaled scoring techniques, and few-shot prompt construction for reliable model behavior.



Jiacun Wang joined Monmouth University in 2004 and is currently a professor of computer science and software engineering. His research interests include machine learning, formal methods, discrete event systems, and software engineering. He has published four books and more than 300 papers. He is the founding EiC of International Journal of Artificial Intelligence and Green Manufacturing, as well as an Associate Editor of IEEE Transactions on Systems, Man, and Cybernetics: Systems, and IEEE/CAA Journal of Automatica Sinica. He has served as general chair and program chair in several international conferences. He is the recipient of 2025 Monmouth University “Distinguished Scholar” Award.