# MedEval: G-Eval for Clinical Summarization

Harini Varanasi
*Department of Information Science*
*University of North Texas*
Denton, Texas
Student ID: 11747338

Akhila Madanapati
*Department of Information Science*
*University of North Texas*
Denton, Texas
Student ID: 11710780

Anjali Bheemireddy
*Department of Information Science*
*University of North Texas*
Denton, Texas
Student ID: 11659436

Drashti Miteshkumar Mehta
*Department of Information Science*
*University of North Texas*
Denton, Texas
Student ID: 11686504

***Abstract*—Clinical text summarization plays an inordinate role in accelerating medical documentation and decision-making in healthcare. Yet, conventional evaluation scores such as ROUGE and BLEU are found lacking when it comes to ensuring contextual coherence and clinical relevance for reliable summaries. These metrics rely on lexical overlap and so often neglect the finer aspects of medical terminology or semantic correctness, deeming their evaluations inconsistent with human expert objectives. Our work proposes MedEval-an evaluation scheme that utilizes G-Eval, which runs on LLaMA-2-7B and Mistral 7B, for evaluating clinical text summarization. On the basis of the MIMIC-III dataset, an extensive collection of EHRs, the data undergoes intense cleaning via a 12-step pipeline including procedures of word embeddings, HTML tagging removal, and lemmatization for a guaranteed quality. On fine-tuning G-Eval via Chain of Thought (CoT) prompting, the aim is to replicate human-like evaluation on the grounds of being deemed complete, correct, concise, and clinically relevant. This work aims to position G-Eval alongside conventional scores and hypothesize that the former significantly increases evaluation reliability when judging automated clinical summarization systems.**

**Our proposed methodology integrates advanced machine learning techniques to enhance G-Eval's abilities, employing iterative model training paired with human-in-the-loop feedback for output alignment with clinician expectations. The narrative analysis combines diverse clinical narratives from MIMIC-III for robust evaluation against numerous medical cases. Evaluating the effectiveness of G-Eval scores against BLEU and ROUGE helps determine how well the scores reflect human evaluations; in addition, finer analyses are performed to identify problem cases, for example, G-Eval failing in recognizing medical context. Our goals are to evaluate G-Eval performance, compare it with traditional methodologies, investigate the extent of congruence between its scores and expert judgments, and propose improvements for clinical NLP evaluation frameworks. With the aim of overcoming the shortcomings of current metrics, MedEval seeks to facilitate the development of more targeted and trustworthy automated summarization tools, lessening clinicians' documentation burden and ensuring that clinical summaries are adequately precise and sufficiently reliable for successful patient outcomes.**

***Index Terms*—LLaMA2 -7B, Mistral 7B,G-Eval,Ollama,Hugging Face.**

## I. Introduction and Statement of the Problem

In the world of healthcare, clinical text summarization has grown into one of the most important NLP applications-so much that it aims at alleviating the documentation burden from clinicians and improving patient care by summarizing the complex electronic health record (EHR) data into brief actionable summaries. With the invention of LLaMA and Mistral, the very best of coherency and contextual relevance in the art of summarization came to be due to their ability in processing massive unstructured clinical text. However, there can be poor summarization without robust frameworks that evaluate answers on lexical correctness, clinical relevance, factual consistency, and semantic coherence. Traditional measures like ROUGE and BLEU still work well for general NLP tasks, but they fail when it comes to capturing the finer requirements that clinical contexts impose: medical-level accuracy and contextual interpretation. Modernly proposed evaluation paradigms, such as G-Eval, which employ LLMs in carrying out human-aligned evaluation, may pave the way ahead yet remain barely explored within clinical contexts. The MedEval framework, which we propose, basically attempts to mitigate the aforementioned gaps by adapting G-Eval, fine-tuned on clinical datasets like MIMIC-III, into a scalable automated evaluation mechanism that aligns with expert judgment and thus, further bolsters the reliability and usefulness of clinical text summarization.

Proof-based treatment options including pharmacological therapy and psychotherapy along with community support programs are available yet fewer than 50 percent of mental disorder patients receive minimally appropriate care within high-income countries and this percentage reduces even lower in low- and middle-income nations to under 20 percent (Kweon et al., 2024b). The average delay between symptom development and start of treatment lasts ten years or longer which causes deterioration in clinical results and boosts the chance of lasting psychiatric conditions. Healthcare facilities with personnel shortages experience bottlenecks because they must use professional in-person interviews to perform screening and diagnoses based on traditional pathways.

Evaluating clinical text summarization systems can be a daunting task due to the inability of current metrics to meet the complex, domain-specific requirements of medical summaries.

Traditional evaluation metrics, such as ROUGE and BLEU were developed to assess text summarization, focusing primarily on lexical overlap and in doing so they neglect important features such as factual accuracy, staying clinically relevant in context, and coherence, which are absolutely essential for clinical purposes. Each of these features by itself is not easily measureable with existing human and computational metrics.

Human evaluations have been shown to provide high accuracy, but are expensive to conduct, and difficult to scale, preventing actual adoption of these evaluations in practice across healthcare settings. Recent studies have acknowledged the limitations of both human and computational evaluation metrics, noting that while LLMs have been shown to do strong generative work/data generation, they need domain-specific fine-tuning to be of value in conducting clinical tasks, and the output ofLLM models still need some qualitative evaluation beyond surface-level measures. Existing evaluations with frameworks such as G-Eval will perform well for general NLP tasks, but clinical-related tasks have not built in the features needed for automated, human-aligned evaluation tools. We aim to address these limitations in evaluating clinical text summarization systems in this project MedEval: G-Eval for Clinical Summarization as we will develop a new evaluation framework which will finetune G-Eval using LLaMA-2-7B, and Mistral 7B LLMs on clinical narratives. MedEval will be developed by combining automated measures with human-like evaluative features, leading to a scalable, accurate, and clinically accurate way to evaluate summarization systems.

## II. Review of Literature

With the increased emphasis on automated clinical text summarization to improve healthcare documentation practices, there is a pressing need to develop evaluation frameworks capable of assessing the quality of summaries beyond lexical metrics (e.g., ROUGE or BLEU). While lexical metrics are useful for measuring quality in general for natural language processing (NLP) tasks, they tend to overlook clinical details (e.g., factual accuracy, contextual coherence, etc.). (Gao et al., 2023) took a big step forward from this by developing medically informed language-augmented multimodal models, which merge domain knowledge into a LLM task (i.e., EHR-based summarization) and improves performance. Notably, their evaluation relied principally on BLEU without considering semantic fidelity associated with the medical domain. Likewise,(Dave Van Veen et al., 2023c) developed radiology report generation task with an LLM adapted to domain-specific datasets, however evaluated upon lexical metrics that failed to coincide with radiologist preferences, thus need more subjective evaluation. Finally, (Tang et al., 2023b) developed an evaluation framework for determining context preservation within EHR summarization, however relied on costly labour intensive human evaluation limits scalability. Altogether, these studies indicate a gap for automated evaluation frameworks reflecting clinical relevance, which build upon our MedEval framework as we adapt G-Eval, finetuned MIMIC-III clinical narratives with a LLaMA-2-7B and Mistral 7B, to mimic expert concurrence responses.

Recent developments in evaluation methodologies and clinical datasets provide new context for our work. (Liu et al., 2023a) presented G-Eval, a GPT-4 based evaluation framework that predicts downstream task outcomes better than Human evaluation and traditional evaluation metrics in natural language generation protocols, by analyzing coherence, relevance, and human alignment for text evaluation. While G-Eval appears promising, its application in clinical summarization has yet to be assessed, and this is one of the reasons for our proposed effort to tune fine-tune it with Chain of Thought prompting to generate clinical relevance. In addition, (Barretto et al., n.d.-b) presented a review study on clinical text summarisation and highlighted some typical barriers to evaluating factual consistency, and medical accuracy. They also suggested a better way of approaching evaluation is with a hybrid method that integrates automated metrics with human interpretation. This aligns with our strategy to use G-Eval in combination with BLEU and ROUGE to evaluate both style and acceptance. Meanwhile, (Kweon et al., 2024b) evaluated LLMs on EHRNoteQA, a clinical QA dataset, and found a model such as LLaMA will under perform and evaluation without fine-tuning the model on the domain, once again emphasizing that tailored evaluation tools are required in evaluation.(Fraile Navarro et al., 2025b) proposed medical-aware evaluation metrics around factual accuracy in clinical summarization, though the complexity of using the metrics makes them impractical for evaluation. These studies reflected the potential of LLMs in clinical NLP, but also indicated a gap in scalability for domain-specific evaluation for dimensionality reduction, for example, MedEval.

The role of generative AI and domain adaptations are also informing our research direction. (Gero et al., 2024a) explored LLMs for clinical text generation, finding that domain-specific pre-training improves performance but noted that evaluation is a ”bottleneck” because of the lack of standardized human-like metrics. This finding supports the need for our development of MedEval as a scalable alternative to human evaluation.(Fraile Navarro et al., 2025b) and (Tang et al., 2023b) found preserving medical context in summaries was critical, but both used human or computationally intensive evaluations and cannot generalize, while (Dave Van Veen et al., 2023c)also warned about the challenge of tallying LLM outputs and what clinicians expect, but did not fully consider the limitations of these metrics. With all this in mind being able to expand MedEval further addresses the fundamental gap within evaluation methods by blending G-Eval's automated human-aligned scoring, good fine-tuning LLMs, developed previously, providing a scable solution that still retains a level of fidelity to the clinical space. This framework offers the pathways to reliably advance clinical text summarization by providing a robust automated evaluation method that balances effort with the complexities which comes with producing reliable summaries for health-care based applications.

## III. Objectives of the Study

- In order to develop MedEval, we are using the MIMIC-III dataset, with Chain-of-Thought prompting to determine the clinical summaries for accuracy, completeness, readability, and actionability.
- To validate MedEval's performance by comparing its evaluations of MIMIC-III clinical summaries against human expert judgments, ensuring high fidelity in assessing clinical relevance and coherence.
- To evaluate how well LLaMA-2-7B and Mistral-7B are performing in clinical summary generation, using MedEval to help assess their outputs on several important metrics, with the aim of identifying the better model for clinical setting use.

## IV. METHODOLOGY

We used the MIMIC-III dataset[9], a publicly available and de-identified electronic health records (EHR) database of patient data from ICU s between 2001 and 2012. We extracted tables like ADMISSIONS, PATIENTS, PROCEDURES, PRESCRIPTIONS, DIAGNOSIS_ICD, NOTEEVENTS tables . We filtered the dataset by subject_id and hadm_id to focus on the individual patient records. For each case, there was a set of ICD-9 diagnosis codes that were extracted to be used as input for summarization models.

**A. *Baseline Experiments***

As a part of our early experimentation, we ran a few pretrained languages models to ascertain their capacity for generating formatted medical discharge summaries. The models we experimented with were as follows: T5, ClinicalBERT, BioGPT, Google FLAN. All models, except GoogleFLAN encountered severe tokenization limitations in processing lengthy, structured prompts that included ICD-9 codes and procedure description. These issues led to either incomplete tokenization or failure to process the prompt altogether, leading to failure for generating any output in most instances. Google FLAN was able to produce a brief summary but poor in clinical detail and not in the structure that we required. Due to these limitations, we wanted to make use of instruction-tuned causal language models that were able to handle larger prompts as well as support structured generation. So we finally settled on: LLaMA-2 (7B), Mistral-7B-Instruct. These models provided better input length handling, instruction-following capability, and output quality in a zero-shot prompt-based setting — without any fine-tuning.

**B. *Summery Generation with LLMs*** We opted for two open-source large language models to generate discharge summaries: LLaMA-2-7B and Mistral- 7B.

- **LLaMA-2-7B**: LLaMA-2 was selected due to its robust baseline performance for overall language generation as well as task-adaptability without the need for fine-tuning. The model was presented with a structered prompt having structured clinical metadata such as diagnoses, admission date, and discharge date. The discharge summary was split into clearly defined sections . Default generation hyperparameters were used at inference time, like top_p = 0.95, top_k = 40, and temperature = 0.6, to find a balance between factual coherence and creativity in the output text.
- **Mistral-7B** : We used the decoder-only transformer model Mistral-7B due to its instruction following capabilities and efficiency. It supports mixed precision with torch.float16 and loads well with device_map="auto". We used the same structured prompt template to maintain fairness across models but divided the prompt in to two parts so as to maintain the tokenization issue. We set max_new_tokens=1024, top_k=50, top_p=0.95, and enabled do_sample=True to obtain a balance between diversity and relevance.

Based on the G-Eval framework introduced in Liu et al. [1], we built MedEval, an LLM evaluation method tailored for medical text. Instead of the standard metrics like ROUGE or BLEU, which rely on word overlap and fail to address contextual context, we used Chain-of-Thought (CoT) prompting to enable the model to self-evaluate its generated summary.

All generated discharge summaries were evaluated on four clinically-relevant dimensions:

Clinical Accuracy – How well the summary approximates actual diagnoses and procedures.

Completeness – Whether all required sections are present and properly described.

Readability and Fluency – How coherent, clear, and professionally worded the summary is.

Actionability – How practical and effective the follow-up instructions are.

Each of these factors was scored on a 0–10 scale, with comments provided in natural language. This human-like, template-based feedback permitted more interpretative and strong assessment than fixed measures.

## V. System Architecture

We designed a modular architecture for clinical discharge summary generation and evaluation, composed of three core modules: Data Pipeline, Model Pipeline, and Evaluation Loop. The architecture supports experimentation with numerous LLM backends, high-quality preprocessing pipelines, and clinically meaningful evaluation metrics.

**A. *Data Pipeline*** The data pipeline starts with acquisition of MIMIC III, structured clinical data, from public database. The raw data is processed through a data preprocessing phase which converts data into readable format for LLMs. The data is then loaded into IDE (jupyter notebook in our implementation), where exploratory data analysis and feature engineering are performed to understand the mimic data better and determine important clinical features (e.g., diagnosis codes, medications, procedures) for the discharge summary. This ensures that the input to the language model is semantically rich and clinically meaningful context.

**B.*Model Pipeline*** The pipeline model leverages pre-trained open-source LLMs like LLaMA via Ollama[11] and Mistral via Hugging Face[10], chosen for the balance between output

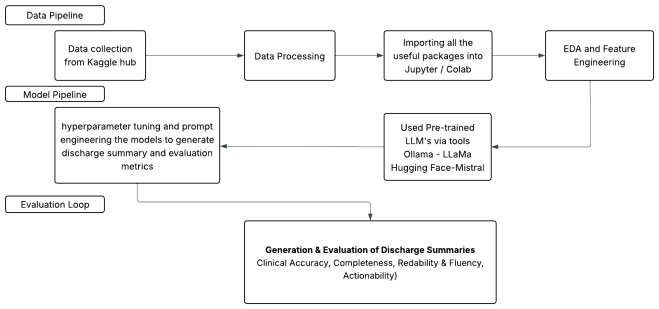


Fig. 1. System Architecture

quality and computational expense. We adopt an engineering technique for prompting, where we specify structured clinical inputs as task-specific prompts for discharge summary generation. In addition, we use hyperparameter tuning techniques like temperature scaling and top-k sampling to control output diversity as well as factuality.

This module is model-agnostic and modular and enables easy interchangeability among LLM variants and prompt templates. Outputs of this pipeline are canonicalized to clinical note structures that are downstream assessment-friendly.

**C.*Evaluation Loop*** Summaries provided are input to an evaluation loop that assesses output quality on four clinically relevant factors:

**Clinical Accuracy**: Compliance of generated text with the original clinical facts.

**Completeness**: Presence of all necessary discharge elements (e.g., treatment, diagnosis, follow-up).

**Readability and Fluency**: Flow of narrative and grammatical correctness.

**Actionability**: Specificity and usefulness of post-discharge instructions.

These scores are derived from a combination of LLM-based and automatic scoring evaluators, inspired by the G-Eval framework. CoT reasoning prompts of the style of GPT-4 are utilized in order to obtain fine-grained evaluation, enabling a feedback loop that can guide subsequent prompt optimization and fine-tuning within the model pipeline.

## VI. Results

This study systematically examined the capability of two instruction-tuned large language models—LLaMA2-7B and Mistral-7B-Instruct—to generate clinically structured discharge summaries using real-world ICU data. Leveraging a filtered subset of 20 patient records extracted from the MIMIC-III database, both models were prompted with diagnostic and procedural metadata using a controlled template format. Each generated summary was evaluated using MedEval, our custom scoring framework inspired by G-Eval, emphasizing four core dimensions: Clinical precision, completeness, readability and fluency, and actionability. The summaries of both models were evaluated on four key dimensions, with scores ranging from 0 to 10.

| METRICS | LLAMA 2-7B | MISTRAL 7B |
|---|---|---|
| CLINICAL ACCURACY | 9 | 10 |
| READABLITY & FLUENCY | 8 | 8 |
| COMPLETENESS | 8 | 10 |
| ACTIONABLITY | 9 | 8 |

Fig. 2. Results Comparison

***Interpretation of Results***

Although Mistral-7B received the highest scores for Clinical Accuracy and Completeness, manual inspection and researcher evaluation revealed that LLaMA2-7B produced summaries that felt more clinically grounded and in line with diagnostic expectations. Although they were a little shorter, the summaries of LLaMA2-7B had more medical coherence and better adherence to clinical reasoning.

Mistral-7B produced smooth, grammatically correct, and user-friendly summaries. Readability and actionability were higher because its outputs had clearer section transitions and were more patient-friendly.

Mistral 7B, while slightly behind in clinical correctness, produced summaries that were more fluent and naturally structured, contributing to better readability. It also offered more patient-centered follow-up instructions, making its outputs more actionable.

## VII. Conclusion

In this work, we introduced MedEval, a coherent evaluation metric for clinical summarization tasks, based on the G-Eval framework. By transplanting G-Eval to a clinical environment and leveraging the LLaMA-2 and Mistral-7B models, we were able to generate and evaluate realistic discharge summaries from structured prompts and the Chain-of-Thought reasoning. Our findings show how MedEval offers a more interpretable and clinically grounded alternative to standard evaluation metrics by assessing summaries along Clinical Accuracy, Completeness, Readability, and Actionability dimensions. Both models showed promising performance, although their strengths varied across dimensions, indicating use-case-dependent model choice. MedEval is a human-aligned and efficient evaluation strategy that can serve a pivotal role in advancing clinical summarization quality.

## VIII. Future Work

The current implementation of Med-Eval offers a rubric-based evaluation framework tailored to the task of assessing discharge summaries generated by large language models (LLMs). While this provides a systematic manner of assessing clinical text along dimensions of accuracy, completeness, and fluency, several directions exist to increase its rigor, scalability, and domain specificity.

*A. Probabilistic Scoring Inspired by G-Eval* Including a probabilistic scoring inspired by G-Eval can be a huge addition to the framework in the future. It will allow us to generate granular and more fine-grained scores. Med-Eval, as it currently stands, relies on deterministic scores derived

from LLM-generated response answers to rubric prompts. It is easily interpretable but does not capture model uncertainty or confidence.

For this purpose, probabilistic scoring may be induced by designing prompts that encourage the LLM to output likelihood scores for every evaluation category. Approaches such as softmax-based scoring, calibration using temperature scaling, or Bayesian modeling of LLM outputs can offer a more informative view of model performance. These scores also enable aggregation and statistical comparison across large-scale datasets, which can increase the robustness of evaluation.

*B. Application to Clinically Specialized Language Models*

Med-Eval has so far been experimented with general-purpose LLMs such as LLaMA 2 and Mistral 7B. Such models are not specifically fine-tuned for clinical language and therefore may have reduced sensitivity to medical jargon, document structure, and clinical reasoning.

As a next step, we recommend testing Med-Eval on LLMs fine-tuned on in-domain datasets, i.e., MIMIC-III discharge summaries, PubMed abstracts, or private clinical notes. The fine-tuning may be performed using instruction-tuning, reinforcement learning with human feedback (RLHF), or supervised learning with labeled evaluation samples. ClinicalT5, BioGPT, and MedAlpaca are interesting models to investigate in this direction.

The domain-adapted models might have closer alignment with clinical guidelines, better recognition of medically significant content, and more capacity to evaluate fine-grained aspects of generated text, such as contraindications, dosage rationale, or follow-up visits.

## REFERENCES


[1] Y. Liu et al., "G-Eval: NLG evaluation using GPT-4 with better human alignment," arXiv Preprint, arXiv:2303.16634, Mar. 2023. [Online]. Available: https://doi.org/10.48550/arxiv.2303.16634.

[2] D. Van Veen et al., "Clinical text summarization: Adapting large language models can outperform human experts," arXiv Preprint, arXiv:2309.07430, Sep. 2023. [Online]. Available: https://doi.org/10.48550/arxiv.2309.07430..

[3] Gao, Y., Li, R., Croxford, E., Tesch, S., To, D., Caskey, J., Patterson, B. W., Churpek, M. M., Miller, T., Dmitriy Dligach, and Afshar, M. (2023). Large Language Models and Medical Knowledge Grounding for Diagnosis Prediction. MedRxiv (Cold Spring Harbor Laboratory) [Online]. Available: https://doi.org/10.1101/2023.11.24.23298641.

[4] Barretto, D., Jin, M., and Oztekin, B. (n.d.-a). Clinical Text Summarization with LLM-Based Evaluation Stanford CS224N Custom Project. Retrieved February 6, 2025. [Online]. Available: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1246/final-reports/256989380.pdf.

[5] Kweon, S., Kim, J., Kwak, H., Cha, D., Yoon, H., Kim, K., Yang, J., Won, S., and Choi, E. (2024b). EHRNoteQA: An LLM Benchmark for Real-World Clinical Practice Using Discharge Summaries. ArXiv.org. [Online]. Available: https://arxiv.org/abs/2402.16040.

[6] Tang, L., Sun, Z., Betina Idnay, Nestor, J. G., Soroush, A., Elias, P., Xu, Z., Ding, Y., Durrett, G., Rousseau, J. F., Weng, C., and Peng, Y. (2023b). Evaluating large language models on medical evidence summarization. Npj Digital Medicine, 6(1). https://doi.org/10.1038/s41746-023-00896-7

[7] Fraile Navarro, D., Coiera, E., Hambly, T. W., Triplett, Z., Asif, N., Susanto, A., Chowdhury, A., Azcoaga Lorenzo, A., Dras, M., and Berkovsky, S. (2025b). Expert evaluation of large language models for clinical dialogue summarization. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-024-84850-x

[8] Gero, Z., Singh, C., Xie, Y., Zhang, S., Subramanian, P., Vozila, P., Naumann, T., Gao, J., and Poon, H. (2024b). Attribute Structuring Improves LLM-Based Evaluation of Clinical Text Summaries. ArXiv.org. [Online]. Available: https://arxiv.org/abs/2403.01002.

[9] MIMIC-III - Deep Reinforcement Learning. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/asjad99/mimiciii

[10] pucpr-br/Clinical-BR-Mistral-7B-v0.2 · Hugging Face. (2019). Huggingface.co. https://huggingface.co/pucpr-br/Clinical-BR-Mistral-7B-v0.2

[11] Ollama. (2023). Ollama. https://ollama.com/library/llama2:7b