MedEval: G-Eval for Clinical Summarization

Harini Varanasi
11747338

Akhila Madanapati
11710780

Anjali Bheemireddy
11659436

Drashti Miteshkumar Mehta
11686504
University of North Texas

February 13, 2025

1 Introduction and Background

In the case of medical text summarization, generated summary accuracy and reliability in medical decision-making become critical. Traditional evaluation tools like ROUGE and BLEU have fallen short in terms of capturing the nuance of human-evaluated text quality. In this work, we aim to explore the performance of G-Eval, an evaluation model with LLaMA-2-7B and Biogpt, in comparing with traditional evaluation tools in terms of clinical text summarization. With a high regard for EHR and clinical documentation summarization, high-quality evaluation methodologies become a necessity. With this project, We aim to assess whether G-Eval generates a sounder and human-conformable evaluation, towards enhancing state-of-the-art automated clinical summarization models.

2 Statement of the Project Problem

Existing evaluation methodologies for clinical text summarization rely almost wholly on lexical overlap-based evaluations such as ROUGE and METEOR, which don't assess contextual cohesion and pertinence. As clinical texts have very specific terminologies and require careful summarization, conventional methodologies cannot accurately capture generated summary effectiveness. G-Eval, powered with LLaMA-2-7B and Biogpt, brings a new direction through use of human-like judgments and deeper language awareness. The most important problem in terms of research is to assess whether G-Eval can outperform conventional evaluation methodologies in assessing clinical text summarization, and in consequence, provide increased evaluation dependability. The work will span a range of phases in the data science lifecycle, including data collection, exploratory data analysis, modeling, visualization, and result interpretation.

1. What are G-Eval's strengths and weaknesses compared to conventional human or computerized evaluation techniques?

2. In what way can G-Eval be tuned to closely mimic human evaluation for clinical text summarization?

3. How can G-Eval most effectively integrate with current NLP evaluation pipelines in a manner that maximizes overall evaluation?

3 Related Work

3.1 *G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment*

G-Eval, a new evaluation tool for NLG, utilizing GPT-4's strong capabilities to best emulate human judgment. Liu et al.'s contribution is a significant one for work in clinical text summarization evaluation. Traditional evaluation methodologies have a flaw in not being capable of emulating nuanced interpretations of humans and, therefore, become less effective for use in a clinical environment. G-Eval addresses such a flaw through a fine-tuned GPT-4 model for evaluation of a text summary. Liu et al. show G-Eval outperforms traditional methodologies such as ROUGE and BLEU in terms of emulating human judgments and providing a more reliable output. G-Eval's performance is contrasted with alternative evaluation methodologies in the work and its utility in improving clinical text summarization, particularly in high-stakes environments in which accuracy and emulating human judgment is important, is highlighted. G-Eval is a powerful tool for improving computerized evaluation tools for use in clinical NLP work[1].

3.2 *EHRNoteQA: An LLM Benchmark for Real-World Clinical Practice Using Discharge Summaries*

EHRNoteQA, a benchmark specifically for testing LLMs with real-life clinical information, such as discharge summaries in EHRs. In its work, the study seeks to tackle the challenge of developing strong NLP models with capabilities for generating and summarizing clinical text in a manner that mirrors medical practice and standards. By providing a corpus of discharge summaries with labels, work in its publication presents a real-life benchmark for testing LLM performance in generating and producing clinical text. In its work, the authors specifically target testing LLMs not only for language accuracy but for medical accuracy and relevance, as well. EHRNoteQA is an important tool for developing NLP for use in practice, with capabilities for testing the quality of generated summaries with expert-curated information and offering an analysis of current LLM-based systems' capabilities in the medical field. In its work, the benchmark aids in developing LLMs for use in practice and harmonizes NLP technology with real-life medical requirements [2].

3.3 *Adapted Large Language Models Can Outperform Medical Experts in Clinical Text Summarization*

The study explores the effectiveness of adapting large language models (LLMs) to clinical text summarization for a range of medical documentation tasks, such as radiology reports, patient questions, progress notes, and doctor-patient dialogue. Through adaptation techniques, the study contrasts the performance of LLMs and medical experts, demonstrating that in the majority of cases, the best adapted models produce summaries that are comparable or superior in completeness, correctness, and conciseness. The study employs both quantitative assessments using NLP metrics and a clinical reader study involving ten physicians. A safety analysis further determines the presence of possible medical risks in AI-generated versus expert-written summaries. The study concludes that LLMs, when properly adapted, can significantly alleviate the documentation burden

on clinicians without sacrificing, and even surpassing, expert-level quality in clinical text summarization [3].

4 Objectives of the Study

The primary objectives of this work are:

1. To evaluate G-Eval for its performance in estimating clinical text summarization.
2. To compare its performance with traditional evaluation methodologies.
3. To investigate whether G-Eval is a better reflection of human evaluation of clinical summaries.
4. To make recommendations for improving evaluation frameworks for use in clinical NLP.

5 Research Design and Methodology

This research capitalizes on the use of the MIMIC-III dataset, applying selected key variables to generate informative clinical summaries to facilitate better medical documentation. Word embeddings during preprocessing enable modeling of sophisticated medical vocabulary to be better processed. The G-Eval system utilizes Large Language Models (LLMs) to generate and score summaries to ensure contextual fidelity and consistency. The text similarity and preservation of content is measured using metrics such as BLEU and ROUGE in comparing reference texts to generated summaries. BLEU is precision-based, while ROUGE is recall-based using measurement of n-gram overlap. By comparing G-Eval's score to that of classic metrics, it is possible to determine its effectiveness in evaluating summary quality and human-judgment consistency. The approach facilitates a strong measurement of clinical text summarization, pointing to G-Eval's potential over classic methods in evaluating medical documentation. The results help to drive forward advances in automatic methods of evaluation, making clinical text summarization more relevant and reliable in healthcare applications.

5.1 Iterative Improvement and Evaluation of Clinical Summaries

This research capitalizes on advanced machine learning approaches to enhance clinical text summarization. Chain of Thought (CoT) prompting is employed in Large Language Models (LLMs) to facilitate more elaborate reasoning and better adherence to clinical expectations, resulting in more accurate and contextual summaries. An iterative testing process is employed to adapt model parameters and prompts to initial estimates of performance, iteratively improving output to higher quality. Probability score calculations, adapted from the underlying paper, also apply to quantitatively determine to what extent created summaries extract key clinical facts. Iterative improvement rounds in future work will be carried out, using actual-world user input in model training and analysis. Feedback mechanisms will be engineered to facilitate model flexibility and enable the summarization process to adapt to actual-world clinical use cases, towards a more resilient and trustworthy machine learning system for medical text summarization.

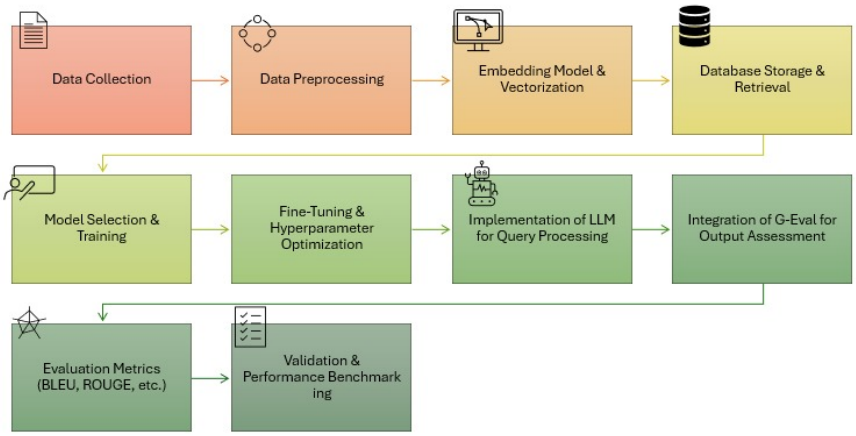

Figure 1: System Architecture

6 Datasets to be Used

The primary dataset for this study is **MIMIC-III**, a widely used clinical data set containing electronic health records. **MIMIC-III** provides a comprehensive source of real-world clinical narratives, making it suitable for evaluating summarization models. The dataset is sufficient for conducting meaningful assessments, as it includes diverse medical cases and notes. If needed, additional clinical text datasets may be incorporated to strengthen the evaluation.

A1 row_id

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	row_id	subject_ic	hadm_id	drg_type	drg_code	descriptio	drg_severi	drg_mortality								
2	1338	10130	156668	HCFA	148	MAJOR SMALL & LARGE BOWEL PROCEDURES WITH COMPLICATIONS, COMORBIDITIES										
3	2188	10114	167957	HCFA	518	PERCUTANEOUS CARDIOVASCULAR PROCEDURES WITHOUT ACUTE MYOCARDIAL INFARCTION, WITHOUT CORONARY										
4	2599	10117	187023	HCFA	185	DENTAL & ORAL DIS EXCEPT EXTRACTIONS & RESTORATIONS AGE >17										
5	2703	10046	133110	HCFA	1	CRANIOTOMY AGE >17 EXCEPT FOR TRAUMA										
6	3020	10011	105331	HCFA	205	DISORDERS OF LIVER EXCEPT MALIGNANCY, CIRRHOSIS, ALCOHOLIC HEPATITIS WITH COMPLICATIONS, COMORBIDIT										
7	4058	10069	146672	HCFA	24	SEIZURE & HEADACHE AGE >17 WITH COMPLICATIONS, COMORBIDITIES										
8	5375	10112	188574	HCFA	138	CARDIAC ARRHYTHMIA & CONDUCTION DISORDERS WITH COMPLICATIONS, COMORBIDITIES										
9	5440	10035	110244	HCFA	5	EXTRACRANIAL VASCULAR PROCEDURES										
10	5635	10111	174739	HCFA	115	PERMANENT CARDIAC PACEMAKER IMPLANT WITH ACUTE MYOCARDIAL INFARCTION, HEART FAILURE OR SHOCK, OR										
11	6612	10013	165520	HCFA	416	SEPTICEMIA AGE >17										
12	7125	10117	105150	HCFA	89	SIMPLE PNEUMONIA & PLEURISY AGE >17 WITH COMPLICATIONS, COMORBIDITIES										
13	11030	10042	148562	HCFA	109	CORONARY BYPASS WITHOUT CARDIAC CATHETER										
14	11337	10032	140372	HCFA	218	LOWER EXTREMITY & HUMERUS PROCEDURES EXCEPT HIP, FOOT, FEMUR AGE>17 COMPLICATIONS, COMORBIDITIES										
15	11539	10061	145203	HCFA	121	CIRCULATORY DISORDERS WITH ACUTE MYOCARDIAL INFARCTION & MAJOR COMPLICATION, DISCHARGED ALIVE										
16	12639	10044	124073	HCFA	10	NERVOUS SYSTEM NEOPLASMS WITH COMPLICATIONS, COMORBIDITIES										
17	13960	10106	133283	HCFA	20	NERVOUS SYSTEM INFECTION EXCEPT VIRAL MENINGITIS										
18	14327	10104	177678	HCFA	320	KIDNEY & URINARY TRACT INFECTIONS AGE >17 WITH COMPLICATIONS, COMORBIDITIES										
19	14743	10056	100375	HCFA	416	SEPTICEMIA AGE >17										
20	15211	10132	197611	HCFA	79	RESPIRATORY INFECTIONS & INFLAMMATIONS AGE >17 WITH COMPLICATIONS, COMORBIDITIES										

Figure 2: Sample Visual Representation of the MIMIC-III dataset

7 Contributions

7.1 Harini Varanasi

Will lead the development of the project, including data preprocessing and model training, ensuring the clinical data in the MIMIC dataset is properly cleaned and standardized efficiently for training the model.

7.2 Akhila Madanapati

Will lead the application and adaptation of the G-Eval framework to assess the quality of generated summaries and will be instrumental in refining the evaluation methodology and performance comparison.

7.3 Anjali Bheemireddy

Will be contributing to integrate the LLM into the project, overseeing the model training process, and tuning model parameters to improve text summarization quality.

7.4 Drashti Miteshkumar Mehta

Will lead the statistical analysis of model outputs, comparing them to standard measures and providing insights to guide iterative improvements to training and evaluation stages.

8 Conclusion

This project seeks to explore the potential of G-Eval in improving the evaluation of clinical text summarization. By comparing it with conventional evaluation methods, the study will determine whether G-Eval provides a more accurate and reliable assessment of clinical summaries. The findings will be valuable for researchers and practitioners in healthcare NLP, paving the way for better summarization evaluation techniques.

9 References

[1] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment," Mar. 2023. Available: `https://doi.org/10.48550/arxiv.2303.16634`

[2] S. Kweon *et al.*, "EHRNoteQA: An LLM Benchmark for Real-World Clinical Practice Using Discharge Summaries," *arXiv.org*, 2024. Available: `https://arxiv.org/abs/2402.16040`

[3] Dave Van Veen et al., "Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts," arXiv (Cornell University), Sep. 2023, `https://doi.org/10.48550/arxiv.2309.07430`