Abstract Control Number: 17116
Resident
Datta

Optimising Large Language Model (LLM) augmented Speech Recognition for Reporting in Radiology: Achieving Near-Zero Error Rates through advanced Prompt Engineering

Purpose

For radiologists, the daunting task of dictating an extensive report combined with the painstaking process of editing the dictations to remove different kinds of errors can consume a substantial amount of their valuable time. This, in addition to their increasing workload, can significantly amplify physician burnout – a pressing issue that plagues radiologists worldwide.

The inevitable delays in reporting a scan, caused by the escalating workloads and the unavailability of affordable proprietary technologies globally which can help radiologists, have made it imperative to harness scalable technologies to enhance the reporting efficiency of radiologists. This is especially true, given that the turnaround time (TAT) for radiology reports can vary drastically, stretching from a few hours for admitted patients to several days in many health systems. [1]

Artificial Intelligence (AI) augmented speech recognition can assist in generating more accurate reports and decrease dictation errors, thereby reducing the time needed for post-dictation error detection. However, there are challenges to the broad application of general speech recognition softwares in the healthcare setup, since radiology reporting contains significant out-of-vocabulary words which traditional open-source softwares fail to recognise correctly. Also, there are significant accent differences among radiologists globally.

Natural Language Processing (NLP) models have achieved benchmarking performance for speech-to-text generation tasks for various applications. However, from our practical experience, most generally available NLP models are unable to achieve good performance in radiology dictation due to the complex medical terminologies used in radiology reports.

The performance of NLPs can be enhanced by augmenting their outputs by Large Language Models (LLMs). AI has been introduced in radiology workflows to improve healthcare and

reduce costs by shortening of the reading time, improved diagnostic accuracy, etc. Automated radiology report generation has recently garnered attention and novel techniques have been explored. In one of our own recent works, we have demonstrated that LLMs significantly improve the output of NLP Models not specifically made for interpreting radiological dictations. However, there is still scope of improvement of LLMs through various methods of prompt manipulation - a term known as prompt engineering.

Limited studies have been conducted only recently to demonstrate the significance of prompt engineering in radiology report generation. PromptRRG [2] was proposed as a method utilizing prompt learning for activating a pretrained model and incorporating prior knowledge, while categorizing them on the basis of varying levels of knowledge: common, domain-specific and disease-enriched prompts. Also, the development of prompt engineering to substantially assist healthcare NLP applications such as question-answering systems, text summarization, and machine translation has also been recently explored. [3]

However, to our knowledge, no studies have been published or presented which have evaluated the incremental effect of staged prompt engineering on the reduction of errors in radiology dictations, especially over any state-of-the-art (SOTA) NLP-LLM pipeline created using LangChain and prompt chaining. Prompt chaining can allow us to accomplish a complex task by passing multiple smaller and simpler prompts. Understanding incremental improvements through prompt engineering is crucial to achieve zero-error rates in automatic speech recognition in radiology.

Methods

Dataset Collection and Preparation

HRCT Chest Reports: A total of 200 HRCT chest reports were randomly selected for this study. These reports were anonymized to ensure patient confidentiality.

Collaborator Involvement: Dictations of the selected reports were obtained from ten individual collaborators. These collaborators represented a diverse group within the medical community, including radiologists, radiologist trainees, radiation oncologists, medical students and non-radiology physicians.

Diversity in Accents: In an attempt to account for the rich linguistic diversity of radiologists, dictations were sourced from collaborators with at least five distinct native accents viz. Hindi, Bengali, Telugu, Tamil and Malayalam.

Dictation Duration: The dictations, on average, spanned a duration of one minute.

Audio File Processing: Given the disparate audio formats in which the dictations were originally received, a standardization process was undertaken. All audio files were converted to the .wav format, ensuring compatibility with the majority of the Natural Language Processing (NLP) models we aimed to evaluate.

Baseline Performance of NLP-LLM Speech Recognition pipeline

The 200 audio dictations and their corresponding signed final text reports were utilized as inputs and ground truth, respectively. A sequential **NLP (WhisperV2 - Large) and LLM (LLaMA v2 7B) pipeline** (which demonstrated SOTA performance in our previous study) was used to process the audio dictations to generate a final text report. It was compared to the ground truth reports to check for errors and baseline performance was recorded.

Layer 1: NLP Model

First, the audio clips are converted into .wav format and then fed to the **Whisper V2-large model** (henceforth termed **Whisper**). Whisper is an automatic speech recognition (ASR) system trained on 680,000 hours of multilingual and multitask supervised data collected from the web. The architecture of Whisper is a simple end-to-end approach, implemented as an encoder-decoder Transformer. Input audio is split into 30-second chunks, converted into a log-Mel spectrogram, and then passed into an encoder. A decoder is trained to predict the corresponding text caption, intermixed with special tokens that direct the single model to perform tasks such as language identification, phrase-level timestamps, multilingual speech transcription, and to-English speech translation.

Layer 2: LLM for processing

The generated text outputs are subsequently fed to a second layer of LLM i.e., LLaMA (Large Language Model Meta AI). The **LLAMA v2 7B model** used here is trained using 7 billion parameters and on a data set with 2 trillion tokens. It is an auto-regressive language model that uses an optimized transformer architecture enabled with supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align to human preferences for helpfulness and safety.

These models were chosen as they demonstrated SOTA performance for the automated speech to report generation. The baseline performance of this sequential pipeline was recorded.

Sequential Prompt Engineering

Selection of N-grams

"N-grams" are continuous sequences of 'n' items in a given text or speech data. For instance, in the sentence "There is no mediastinal lymphadenopathy", the 2-grams (or bigrams) are "There is", "is no", "no mediastinal", and "mediastinal lymphadenopathy". N-grams play a pivotal role in natural language processing and computational linguistics, serving as foundational elements for various models. When applied to prompt engineering, especially in language models, n-grams can be instrumental. They allow for the analysis of context, aiding in the prediction of subsequent words based on prior sequences. By understanding and manipulating n-grams, prompt engineers can craft prompts that guide models more effectively towards contextually relevant responses.

For our study, N-grams that appeared more than 50 times in a set of randomly selected 250 Chest CT reports were extracted and used for prompt engineering in three sequential stages by adding the unigram, bi-gram and tri-gram terms in the context and prompts.

Staged Prompt Engineering

Large language models (LLMs) can be utilized more efficiently and economically through prompt engineering. This approach focuses on meticulous design and refinement of the prompts given to these models. By aiding LLMs in swiftly adjusting to task structures and retrieving pertinent pre-trained knowledge that aligns with the domain-specific requirements of the task, it offers an optimal means to enhance the precision and applicability of the text generated for radiology reports.

For our study, we have used the following prompt and chain these prompts in successive stages to get a better result:

Prompt: “Given a vocal transcript of a medical report detailing the findings of a CT scan, identify and correct any inaccuracies, inconsistencies, or ambiguities to ensure the highest level of clinical accuracy. The transcript may contain medical terminology, abbreviations, and complex sentence structures commonly found in radiology reports. Your goal is to produce a corrected version that maintains the original meaning while adhering to medical standards and best practices.”

Subsequent to this prompt, the following engineering stages were used incorporate additional prompts:

Stage 1:
Using the prompt templates available in LangChain, we include the list of extracted unigrams, and pass them to the model along with our prompt.

Stage 2:
Using the prompt templates available in LangChain, we include the list of extracted unigrams & bi-grams, and pass them to the model along with our prompt.

Stage 3:
Using the prompt templates available in LangChain, we include the list of extracted unigrams & bi-grams & tri-grams, and pass them to the model along with our prompt.

The following **Figure 1** shows the Top 10 common unigrams, bi-grams and tri-grams among the n-grams added in context during our prompt engineering process.

Figure 1: Top 10 common n-grams used for Prompt Engineering

Results

Performance Evaluation Metrics

To assess the efficacy of the applied NLP and LLM models, we employed a series of commonly used metrics:

I. **Word Error Rate (WER):** This metric scrutinizes the predicted output against the reference transcript on a word-to-word basis, identifying discrepancies. It is particularly useful when evaluating the accuracy of speech-to-text transcriptions.

II. **Sentence Error Rate (SER):** SER quantifies the percentage of sentences that contain at least one error in word transcription.

III. **Match Error Rate (MER):** MER provides an estimate of the likelihood of an identified match being erroneous. It's notable that MER values are always less than or equivalent to WER values.

IV. **Average Levenshtein Distance (LD)**: This string metric gauges the dissimilarity between two sentence sequences, offering a broader perspective on the accuracy of the transcriptions.

V. **Semantic Similarity (SS):** SS measures the degree to which two pieces of text carry the same meaning. This metric is especially important in tasks where the exact phrasing might differ, but the underlying message or intent remains consistent. High semantic similarity scores indicate that the compared texts are semantically close, even if they have lexical or syntactic differences.

Experimental Outcomes

The outcomes for each experiment is enumerated in Table 1. The baseline Whisper-LLAMA v2 7B model showed performance metrics of **WER=0.0306; SER=0.019; MER=0.0164; LD=0.65 and SS=0.898.**

The application of prompt engineering resulted in significant improvements across all metrics. After all the three stages of prompt engineering, we achieved a **WER=0.0183, SER=0.013, MER=0.0094, LD=0.38, and SS=0.977**, with no clinically significant errors in any of our reports (as confirmed by a consensus of two radiologists), indicating near-perfect accuracy in report generation.

Table 1: Performance Evaluation of the NLP-LLM Pipeline through various stages of Prompt Engineering

Model Development Stage		WER	SER	MER	LD	SS
Baseline		0.0312511	0.019	0.0164	0.65	0.8984793
Prompt Engineering	Stage 1	0.0268426	0.0171	0.014924	0.5655	0.9262239
	Stage 2	0.0234294	0.0145	0.0132823	0.491985	0.9572427
	Stage 3	**0.0190170**	**0.0127**	**0.0094304**	**0.383748**	**0.9868800**

Figure 2: Comparison of Performances through different Prompt Engineering Stages

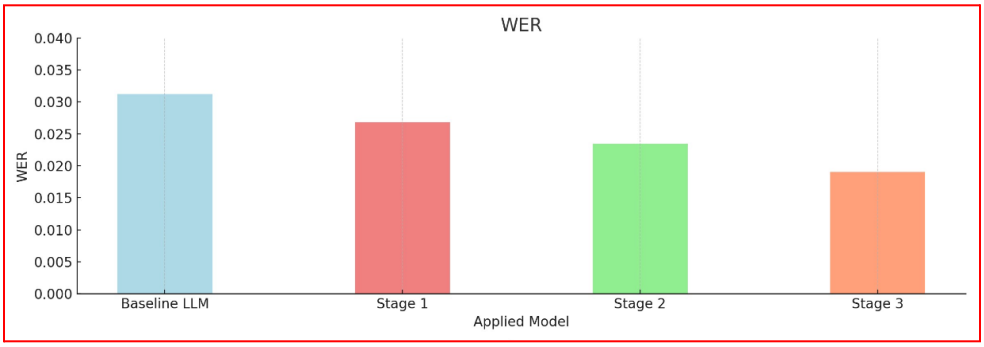


Figure 2 (Continued):

Comparison of Performances through different Prompt Engineering Stages

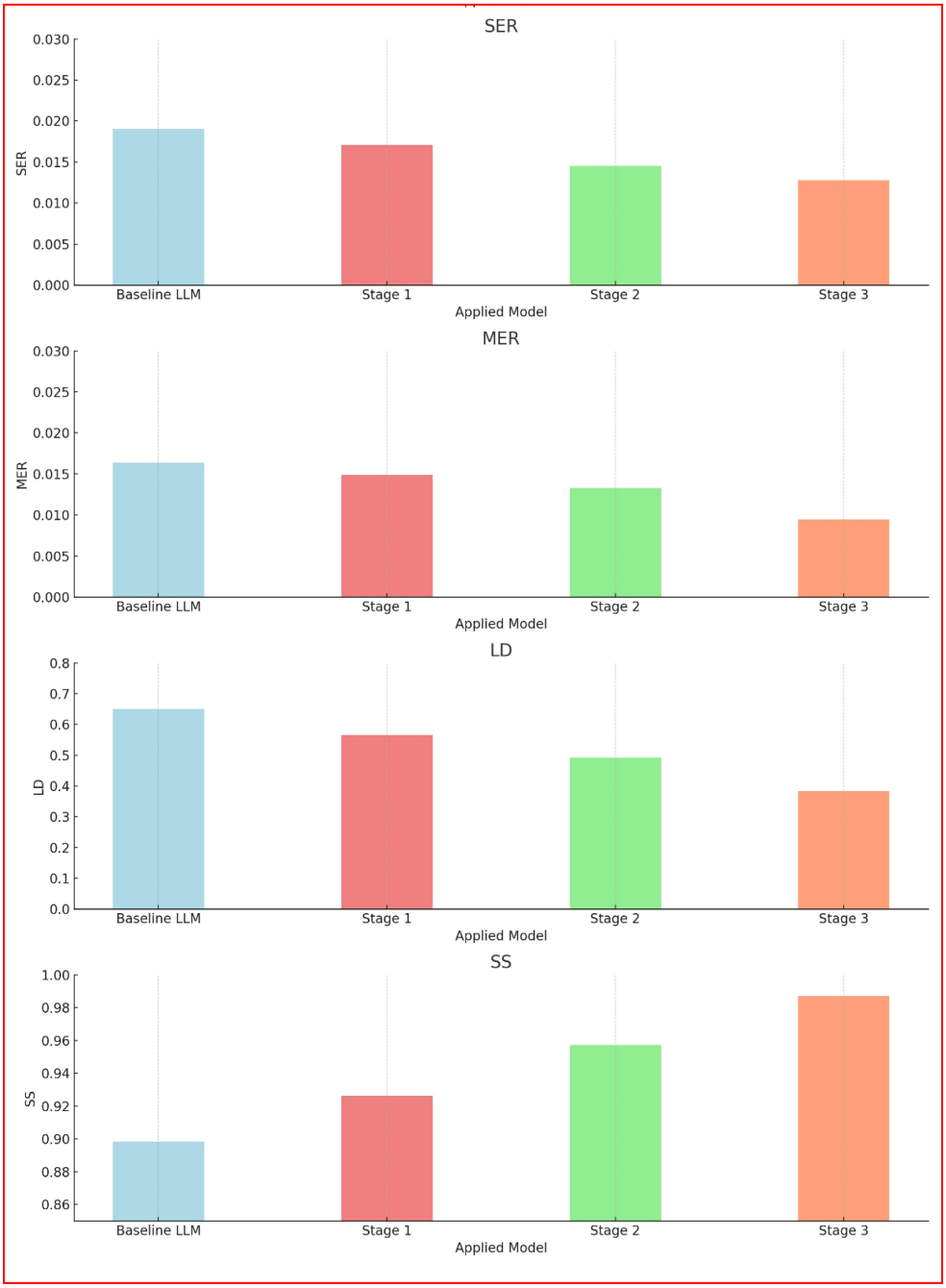


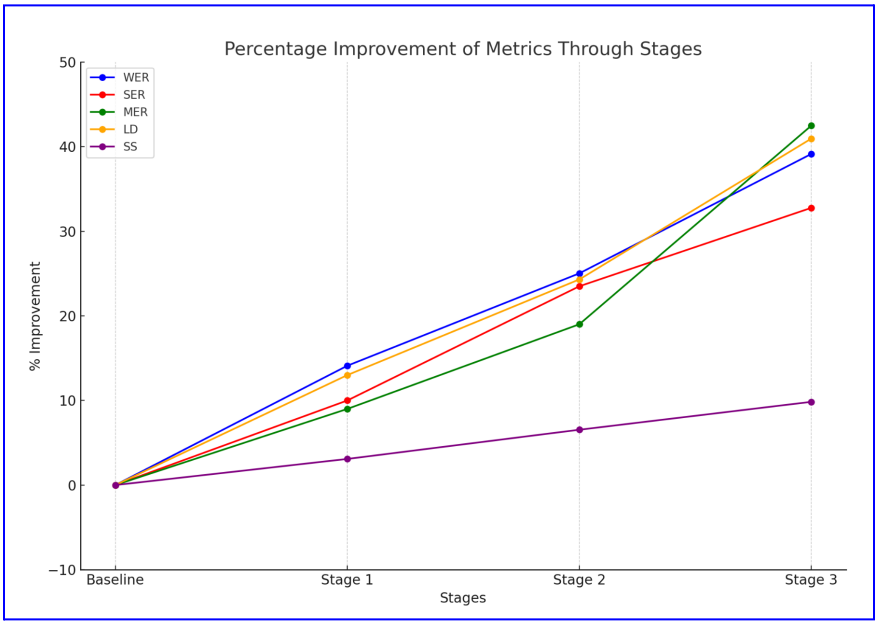
Table 1 and **Figure 2** demonstrate that prompt engineering stages are definitely enhancing the text quality as the values of errors like WER, SER, MER, and LD are decreasing while SS value which shows similarity is increasing.

From the baseline model to the third stage prompt engineering, there is significant improvement across all metrics as demonstrated in **Table 2** and **Figure 3.**

Table 2: Improvement in performance (%) of our NLP-LLM Pipeline compared to baseline metrics after various stages of Prompt Engineering

Stage	Reduction of WER (%)	Reduction of SER (%)	Reduction of MER (%)	Reduction of LD (%)	Increase in SS (%)
Stage 1	14.10646	10	9	13	3.08795
Stage 2	25.0283	23.5	19.01	24.31	6.54032
Stage 3	**39.1477**	**32.77368**	**42.4971**	**40.9618**	**9.83892**

Figure 3: Staged improvement in performance metrics through Prompt Engineering



Conclusion

Our study establishes that prompt engineering can significantly improve the performance of LLM augmented Speech Recognition pipelines for transforming radiological dictations into final reports. In our study, we found that the Whisper and LLAMA v2 7B Speech Recognition pipeline, already recognized for its state-of-the-art performance, showed significant improvement across all metrics through advanced prompt engineering methods. This brings us closer to generating radiological reports that are nearly errorless and free from clinically significant mistakes.

We show that incorporating common n-grams in context can significantly reduce the errors in radiology reports and this is one of the first attempts to demonstrate the utility of the same in radiology report generation.

The augmented Speech Recognition pipelines after advanced prompt engineering significantly improves the accuracy of radiological dictations regardless of a radiologist's accent, minimizes the need for post-dictation error correction and thereby can significantly improve reporting turnaround times.

References

1. Canon CL, Chick JF, DeQuesada I, Gunderman RB, Hoven N, Prosper AE. Physician burnout in radiology: Perspectives from the field. Am J Roentgenol. 2022;218(2):370-4.
2. Wang J, Zhu L, Bhalerao A, He Y. Can Prompt Learning Benefit Radiology Report Generation? arXiv preprint arXiv:2308.16269. 2023.
3. Wang J, Shi E, Yu S, Wu Z, Ma C, Dai H et al. Prompt engineering for healthcare: Methodologies and applications. arXiv preprint arXiv:2304.14670. 2023.