Abstract Control Number: 16936
Resident
Datta

A comparative analysis of Speech Recognition Pipelines augmented by Large Language Models (LLMs) for Radiology Reporting: A 2023 Update

Purpose

Burnout is a global health problem affecting physicians across all medical specialties. Radiologists in particular, experience high burnout due to many factors including increasing workload, inefficient electronic health records, cumbersome hospital policies and excessive call duties. [1]

Increasing workloads often lead to delay in reporting and leveraging technology to improve the efficiency of reporting is crucial to reduce turnaround time (TAT) of radiology reports which can often range from hours to days.

Artificial Intelligence (AI) augmented speech recognition is a promising technology that can assist radiologists in generating more accurate reports and decrease dictation errors, thereby reducing the time needed for post-dictation error detection and automatic report correction. 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). In fact, AI has been introduced in radiology workflows to improve healthcare and reduce costs by shortening of the reading time, reduction of dose and contrast agents, earlier detection of disease and improved diagnostic accuracy. Automated radiology report generation has recently garnered attention and novel techniques have been explored. A Multimodal Recursive model with Contrastive Learning (MRCL) by incorporating both visual and

semantic features for generating “Impression” and “Findings” of radiology reports used a contrastive pre-training method to improvise the expressiveness of both visual and textual representations. [2] Few other recent works have explored modern transformer-based encoder-decoder models [3] and multi-modal transformer networks [4]. However, to our knowledge, there are only limited studies utilizing a combination of NLP and LLMs for radiology reporting.

In our study, we aim to identify the most efficient end-to-end Dictation-to-Report Pipeline by comparing the performance of a sequential combination of widely accessible (both proprietary and non proprietary) NLP and LLM models for speech recognition and subsequent report generation. In our pipeline the various NLP models first create a speech-to-text output and subsequently, the errors in the output are reduced by processing them through a second layer of an LLM to generate a final report.

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.

Model Details

Overview of the NLP Models (Layer 1 of the Pipeline)

For our study, we initially compared five state-of-the-art Natural Language Processing (NLP) models to transcribe the dictated audio files into cohesive text. Five models were chosen based on their popularity, availability, recognition capabilities and adaptability to various speech patterns.

1. **Whisper**
    - Description: A proprietary Automatic Speech Recognition (ASR) API specialized in high-accuracy speech-to-text tasks. It employs an end-to-end encoder-decoder Transformer architecture.
    - Parameters: Trained on 680,000 hours of multilingual and multitask supervised data.
    - Architecture: Segments audio into 30-second chunks, converting them to log-Mel spectrograms for processing by the encoder-decoder model.
    - Embeddings: Not specified, but likely utilizes spectrogram-based embeddings given its architecture.

2. **DeepSpeech**
    - Description: An open-source, end-to-end trainable character-level deep recurrent neural network. It uses a Connectionist Temporal Classification (CTC) decoder based on a greedy decoding strategy.
    - Parameters: Contains over 120 million parameters.
    - Architecture: Employs the CTC (Connectionist Temporal Classification) decoder.
    - Embeddings: Likely employs character-level embeddings based on its architecture.

3. **Word2Vec**
    - Description: Comes in two architectures—Continuous Bag of Words (CBOW) and Skip-Gram. Designed to predict context words from a center word or the reverse.
    - Parameters: Not specified.
    - Architecture: Utilizes CBOW and Skip-Gram architectures.
    - Embeddings: Uses word embeddings, relying on either the CBOW or Skip-Gram methods.

4. **Spark NLP**
    - Description: An open-source NLP library optimized for executing parallel prediction tasks.
    - Parameters: Not specified.
    - Architecture: Built on the foundations of Apache Spark and Spark ML.
    - Embeddings: Not specified, but likely supports various embedding techniques due to its Spark ML foundation.

5. **AssemblyAI**
    - Description: A specialized speech recognition model that incorporates transformers and convolutional layers. It features modified sparse attention to counter noise.
    - Parameters: Trained on 650,000 hours of curated English audio.
    - Architecture: Leverages transformers and convolutional layers with progressive downsampling and grouped attention modules.
    - Embeddings: Not specified, but likely uses feature-based embeddings due to its architecture.

Overview of the LLMs for Post-Processing (Layer 2 of the Pipeline)

After transcribing the dictated audio files with the NLP models, we further processed the output using four state-of-the-art Large Language Models (LLMs). The following models were specifically chosen due to their popularity, easy availability and recent demonstrations of achieving comparable performance in general and medical domain applications:

1. **LLaMA**
    - ❖ Developer: Meta AI
    - ❖ Description: An open-source large language model designed for a variety of natural language processing tasks.
    - ❖ Variants Used in Study:
        - ➢ **LLaMA (v2 7B):**
            - ■ Parameters: 7 billion
            - ■ Training Data: 2 trillion tokens
        - ➢ **LLaMA (v2 13B):**
            - ■ Parameters: 13 billion
            - ■ Training Data: 2 trillion tokens

2. **Alpaca**
    - ❖ Developer: Stanford University (fine-tuned from Meta's initial LLaMA 7B model)
    - ❖ Description: A specialized adaptation of the LLaMA model, tailored for specific research applications.
    - ❖ Parameters: 7 billion

3. **Falcon-7B:**
    - ❖ Description: A causal decoder-only model engineered for efficient text generation and other single-directional tasks.
    - ❖ Parameters: 7 billion
    - ❖ Training Data: 1,500 billion tokens from RefinedWeb, enhanced with curated corpora.

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.

Initial Evaluation of the NLP Model (Layer 1)

Of the five NLP models in our study, the **Whisper model** outperformed other NLP models, registering metrics of WER = 0.153, SER = 0.095, MER= 0.082 and LD = 3.25.

However, from a radiological perspective, none of the results from any of the NLPs were deemed optimal for a speech-to-text pipeline, given the prevalence of significant errors in the generated reports.

Evaluation of the Improvement with LLMs (Layer 2)

To augment our initial outputs and improve the final reports, each NLP model's output was processed with each of the four LLMs.

As elucidated in **Table 1 and 2** and shown in **Figure 1** below, the introduction of LLM models significantly improved the performance metrics, with significant error reductions reaching 80% across various metrics.

Table 1: Performance evaluation of the different Radiology Report Generation pipelines

Layer 1 (NLP)	Layer 2 (LLMs)	WER	SER	MER	LD
AssemblyAI	Without LLMs	0.184	0.118	0.091	3.52
	LLaMA v2 7B	0.0368	0.0236	0.0182	0.704
	LLaMA v2 13B	0.04824	0.03092	0.02366	0.80726
	Falcon 7B	0.0552	0.0354	0.0273	1.056
	ALPACA	0.062	0.039	0.03	1.19
DeepSpeech	Without LLMs	0.184	0.118	0.091	3.52
	LLaMA v2 7B	0.0402	0.0278	0.0224	0.772
	LLaMA v2 13B	0.05226	0.03686	0.02912	0.76262
	Falcon 7B	0.0603	0.0417	0.0336	1.158
	ALPACA	0.068	0.046	0.037	1.31
SparkNLP	Without LLMs	0.236	0.163	0.131	4.14
	LLaMA v2 7B	0.0472	0.0326	0.0262	0.828
	LLaMA v2 13B	0.06184	0.04262	0.03406	0.77998
	Falcon 7B	0.0708	0.0489	0.0393	1.242
	ALPACA	0.08	0.055	0.044	1.4
Word2vec	Without LLMs	0.251	0.185	0.152	4.41
	LLaMA v2 7B	0.0502	0.037	0.0304	0.882
	LLaMA v2 13B	0.06526	0.04806	0.04048	0.80674
	Falcon 7B	0.0753	0.0555	0.0456	1.323
	ALPACA	0.085	0.062	0.05	1.49
Whisper	Without LLMs	0.153	0.095	0.082	3.25
	LLaMA v2 7B	**0.0306**	**0.019**	**0.0164**	**0.65**
	LLaMA v2 13B	0.04002	0.02446	0.02132	0.8455
	Falcon 7B	0.046	0.0285	0.0246	0.975
	ALPACA	0.052	0.032	0.027	1.09

Figure 1. Comparison of performance metrics of different Dictation-to-Report pipelines

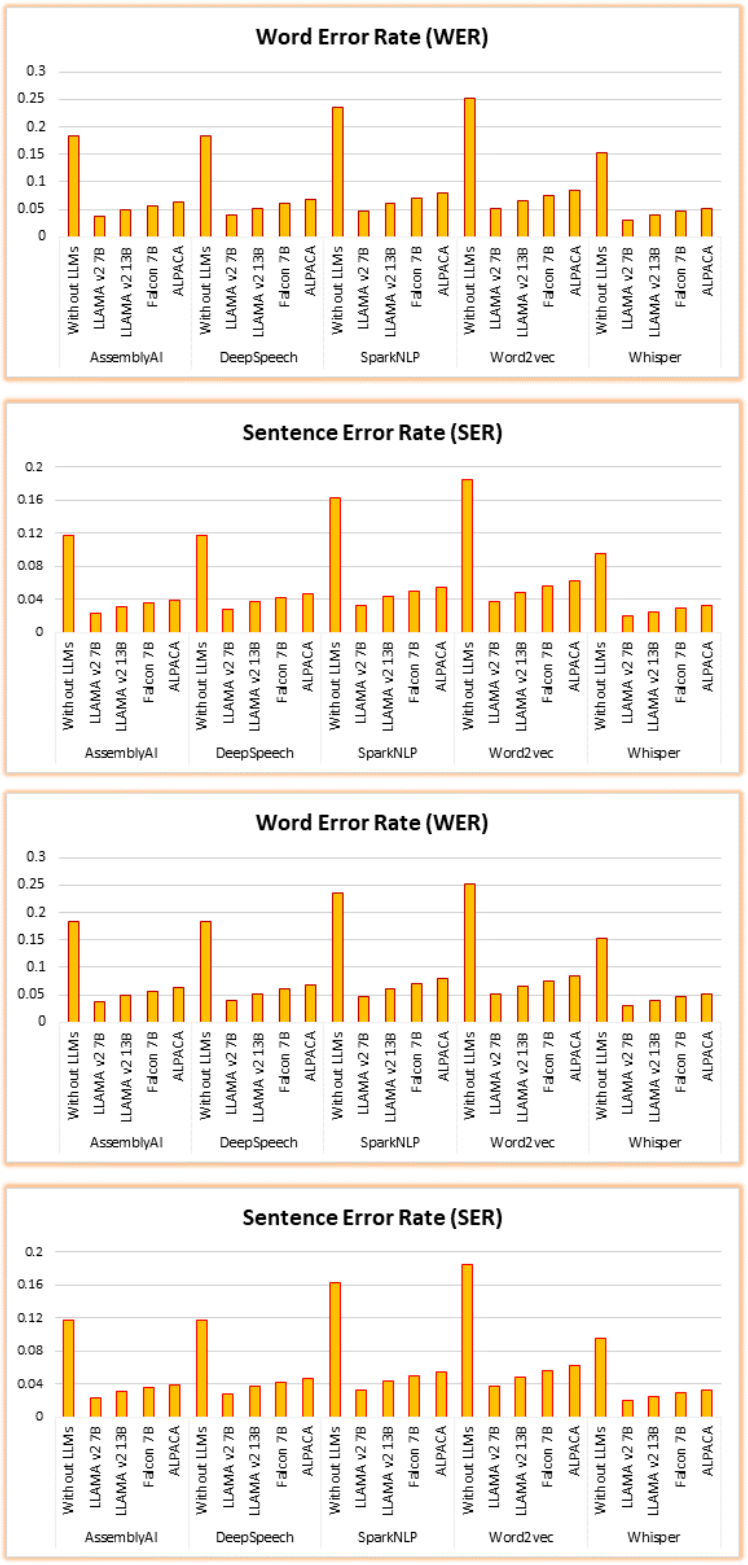
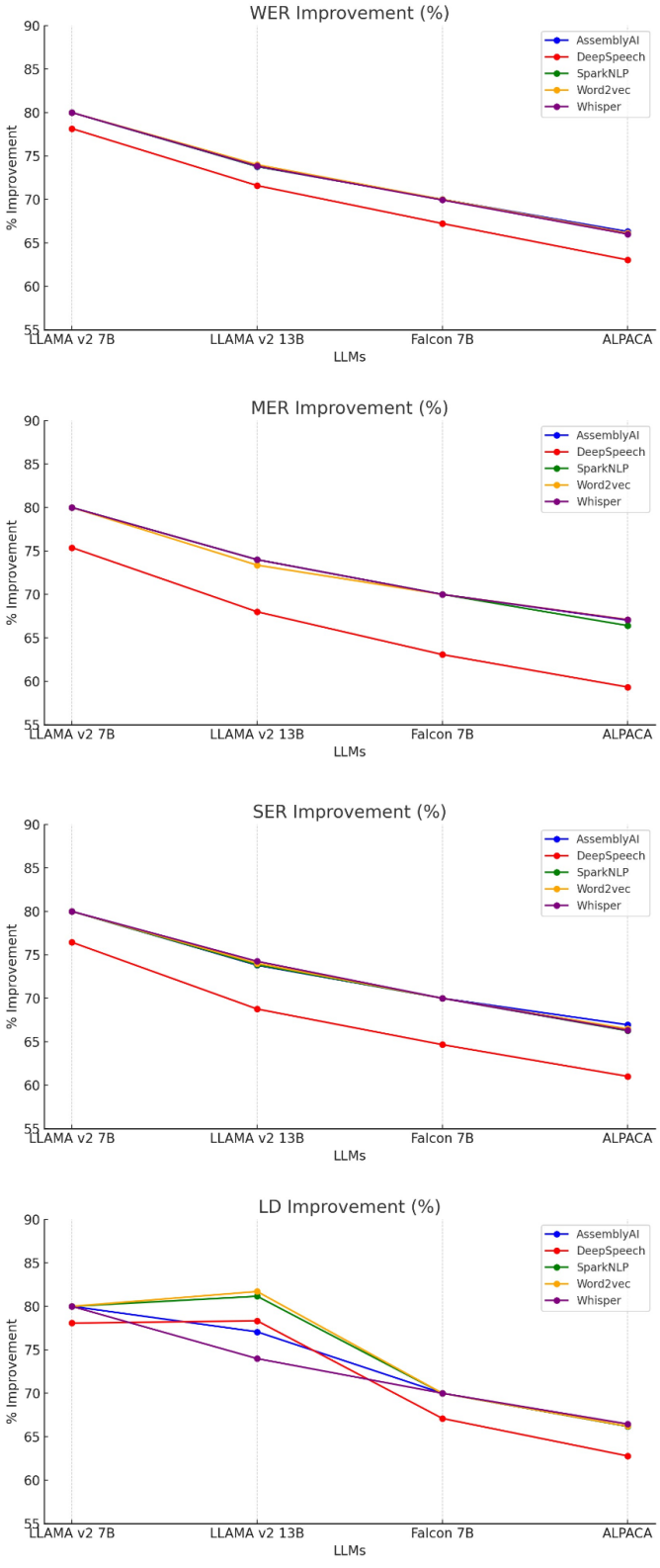


Table 2: Improvement in Performance of various NLP Models (Layer 1) after processing through different LLMs (Layer 2)

NLP Models	LLM	% WER Reduction	% SER Reduction	% MER Reduction	% LD Reduction
AssemblyAI	LLAMA v2 7B	**80**	**80**	**80**	80
AssemblyAI	LLAMA v2 13B	73.782608	73.796610	74	77.066477
AssemblyAI	Falcon 7B	70	70	70	70
AssemblyAI	ALPACA	66.304347	66.949152	67.032967	66.193181
DeepSpeech	LLAMA v2 7B	78.152173	76.440677	75.384615	78.068181
DeepSpeech	LLAMA v2 13B	71.597826	68.762711	68	78.334659
DeepSpeech	Falcon 7B	67.228260	64.661016	63.076923	67.102272
DeepSpeech	ALPACA	63.043478	61.016949	59.340659	62.784090
SparkNLP	LLAMA v2 7B	**80**	**80**	**80**	80
SparkNLP	LLAMA v2 13B	73.796610	73.852760	74	**81.159903**
SparkNLP	Falcon 7B	70	70	70	70
SparkNLP	ALPACA	66.101694	66.257668	66.412213	66.183574
Word2vec	LLAMA v2 7B	**80**	**80**	**80**	80
Word2vec	LLAMA v2 13B	74	74.021621	73.368421	81.706575
Word2vec	Falcon 7B	70	70	70	70
Word2vec	ALPACA	66.135458	66.486486	67.105263	66.213151
Whisper	LLAMA v2 7B	**80**	**80**	**80**	80
Whisper	LLAMA v2 13B	73.843137	74.252631	74	73.984615
Whisper	Falcon 7B	69.934640	70	70	70
Whisper	ALPACA	66.013071	66.315789	67.073170	66.461538

Figure 1: Performance Improvement after processing NLP Model outputs through LLMs



Selection of the best Final Pipeline (LLM combined with NLP Model)

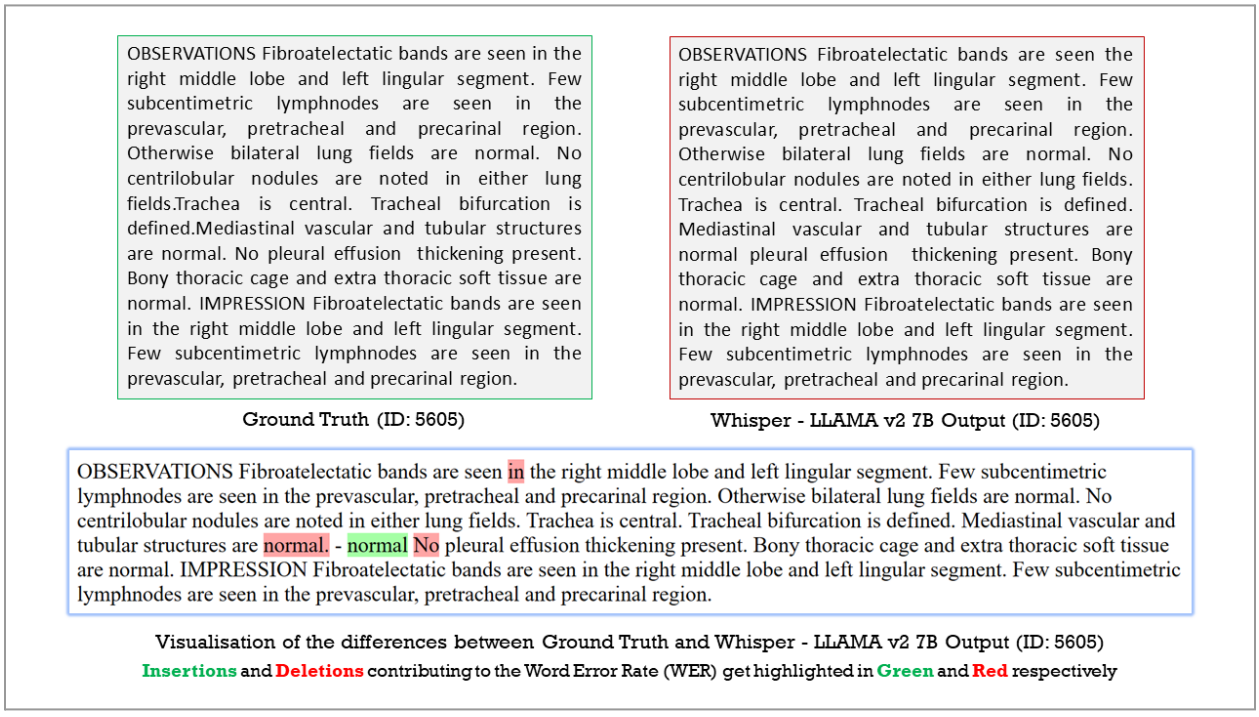
Of the various model combinations assessed, the **combination of the LLaMA v2 7B and Whisper NLP model** yielded the most promising results with WER = 0.0306, SER = 0.019, MER = 0.0164 and LD = 0.65.

Comparison and Evaluation of the Pipelines by Radiologists

A blinded qualitative comparison of the output reports (by grades from 1 to 5) was subsequently done by two radiologists to further evaluate the impact of introducing LLM models in the report generation pipeline using NLP models. It also confirmed the augmented Whisper-LLaMA (v2 7B) pipeline as the best-performing model with a Grade of 4.5.

Figure 2 shows the comparison of one of the Radiology Reports (Ground Truth) written by a radiologist and the generated output of Whisper - LLaMA v2 7B pipeline for speech recognition:

Figure 2. Comparison of Ground Truth Radiology Report and output of the Whisper - LLaMA v2 7B pipeline



Conclusion

The addition of a second layer of LLM for processing the output of a Speech-to-text NLP model significantly reduces the errors and improves the performance of a speech recognition pipeline for radiological dictation.

Among evaluated pipelines, the combination of Whisper and LLaMA v2 7B is the best performing LLM augmented Speech Recognition pipeline for transcribing radiological dictations into final textual reports.

We demonstrate that LLM augmented Speech Recognition Pipelines can significantly improve the accuracy of radiological dictations, minimize requirements for post-dictation error corrections and improve reporting turnaround times.

We also demonstrate that many of our pipelines which are completely built from widely available open-source NLP-LLM combinations show good performance These automatic speech recognition pipelines can be incorporated for radiology reporting and essentially be used globally by radiologists with different accents, without needing to rely on proprietary speech recognition softwares.

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 Feb 8;218(2):370-4
2. Wu X, Li J, Wang J, Qian Q. Multimodal contrastive learning for radiology report generation. J Ambient Intell Humaniz Comput. 2023 Aug;14(8):11185-94
3. Nimalsiri W, Hennayake M, Rathnayake K, Ambegoda TD, Meedeniya D. Automated Radiology Report Generation Using Transformers. In: 2023 3rd International Conference on Advanced Research in Computing (ICARC); 2023 Feb 23. p. 90-95. IEEE.
4. Aksoy N, Ravikumar N, Frangi AF. Radiology report generation using transformers conditioned with non-imaging data. In: Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications; 2023 Apr 10. Vol. 12469. p. 146-154. SPIE.