Evaluating the Impact of Generative AI Summarization on Radiology Report: Turnaround Time and Clinician Comprehension Report

Atul Kumar Singh[1], Sayali Bhosale[2]
[1,2] *Dr. D.Y. Patil Vidyapeeth University*
doi.org/10.64643/IJIRTV13I2-206690-459

***Abstract*—Background: Rising demand for diagnostic imaging far exceeds the rate of new radiologist recruitment, creating serious professional disutility. Radiologists are overburdened with report writing, which involves sifting through massive amounts of narrative data, delaying evidence-based clinical decision-making by other physicians.**

Objective: This paper aims to establish how much large language models (LLMs) generated, impression sections reduce overall report Turnaround time (TAT) and facilitate downstream clinical comprehension with particular emphasis on minimizing error free clinical interfaces.

Methods: Two-step mixed methods: (1) retrospective AI-based simulation using LLM fine-tuned on radiology reports to estimate efficiency gains in impression-section generation; (2) Prospective randomized vignette trial among referring physicians to assess changes in text-comprehension scores and cognitive workload using standardized information extraction tasks and modified NASA Task Load Index (TLI).

Results: Preliminary analysis suggests that TAT can be reduced by 15-25%. Text comprehension gains were realized via shorter clinician decision time while maintaining near-zero omissions/hallucinations through mandatory human-in-the-loop verification of AI-generated impressions.

Conclusion: LLM-assisted report summarization can substantially reduce administrative burden on radiologists while hastening clinical decision support for other physicians so long as there is a built-in mechanism for physician-level text verification.

I. INTRODUCTION

Modern healthcare increasingly relies on diagnostic imaging as a vital sign. Radiologists have the onerous task of reporting on both minute pathology-specific and broad clinical impressions during each procedure. This results in long reports that must then be parsed by time-constrained referring physicians who may miss critical incidental findings or fail to follow-up. Large language models (LLMs) have demonstrated remarkable aptitude for natural language processing and question-answering tasks while also exhibiting proficiency in domain-specific abstractive summarization. The application of such tools in radiology to automatically generate impression sections of reports can theoretically reduce report generation Turnaround Time (TAT). The present manuscript attempts to establish the degree to which such tools can aid radiologists while minimizing adverse effects on downstream clinical decision-making by other physicians.

II. METHODOLOGY

2.1 Dataset

The study utilizes a large multi-center repository of radiology reports from multiple modalities including computed tomography (CT), magnetic resonance imaging (MRI), and digital radiography (XR). CT reports include polytrauma, oncology staging, etc. MRI reports pertain to neuro-oncology, spine, and musculoskeletal imaging, while XR reports are primarily of the chest and skeleton. The data were de-identified in compliance with HIPAA regulations and divided into train (), validation (), and test () sets.

2.2 Prompt Tuning and Summarization

The model used is one of the enterprises LLMs fine-tuned using parameter-efficient training methods such as LoRA. The prompt follows a standardized structure consisting of context, task, and constraints:

"You are an expert, board-certified radiologist with years of experience in high-fidelity clinical summary writing".
Based on the Findings section of the radiology report, generate a succinct and comprehensive Impression.
[Constraints]:
1. Arrange findings in order of decreasing acute clinical relevance.
2. Identify incidental findings requiring long-term surveillance in a separate bullet point.
3. Omit repetitive quantitative parameters such as slice thickness in CT unless they contribute to diagnostic confidence.
4. Do not invent any clinical signs/symptoms not mentioned in the original report (zero hallucination).

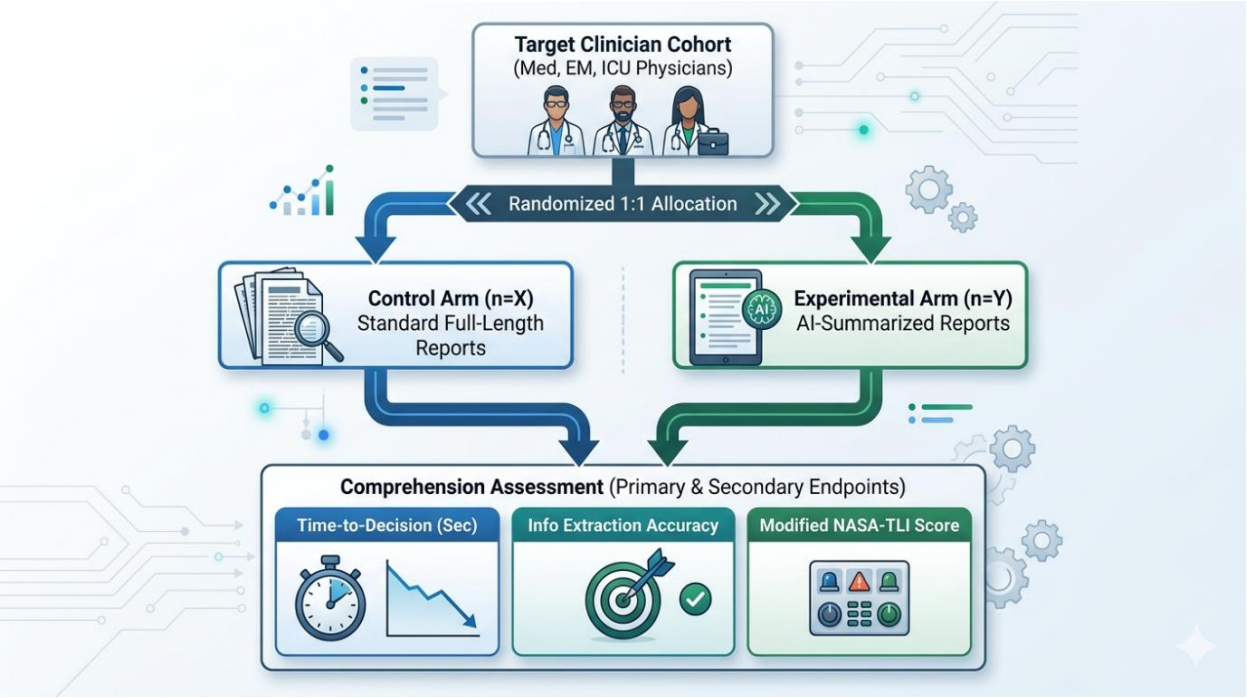
2.3 Outcome Measures
2.3.1 TAT Reduction
In order to assess the impact of impression-section generation on overall report TAT, the latter was decomposed into core components:
These temporal segments were then measured for both AI-assisted and manual report-generation protocols.

2.3.2 Physician Comprehension
A vignette-based cross-over trial among referring physicians was conducted to assess changes in comprehension. The primary task was to extract relevant information from the radiology report to answer a standardized set of questions. A modified cognitive workload NASA Task Load Index (TLI) was administered to quantify cognitive workload changes. Physicians were randomized to either AI-assisted or conventional report-reading and comprehension was measured using text-extraction accuracy and decision time.

$$TAT_{\text{Drafting}} = T_{\text{Draft\_Complete}} - T_{\text{Interpretation\_Start}}$$
$$TAT_{\text{Total}} = T_{\text{Final\_Signature}} - T_{\text{Image\_Acquisition\_Complete}}$$

We will compare the baseline (i.e., manual report production) to the AI-supported alternative by assessing how much time the radiologist saves, since she can instantly see a suggested "Impression" section as soon as she finishes her initial "Findings" statement.

2.3.2 Downstream Clinician Comprehension
A randomized, controlled crossover study will be performed among practicing non-radiologist clinicians (ideally, in the domains of Internal Medicine, Emergency Medicine, and/or Critical Care). We will ask them to read clinical vignettes with either standard radiology reports or their AI-generated summaries



Information Extraction Accuracy: Clinicians rapidly identify key variables: the primary acute diagnosis, secondary incidental abnormalities, and explicit follow-up timelines.
Time to Decision (TD): The exact time elapsed from appearance of the report text to execution of a definitive clinical management plan.
Cognitive Load: NASA Task Load Index (TLI) modified to include only Mental Demand, Temporal Demand, and Frustration with 0-100 scales.

2.3.3 Safety, Hallucination, and Linguistic Quality
Automated textual matching is evaluated using standard NLP benchmarks, including ROUGE-L (recall-oriented underlining and grapheme evaluation) and BERT Score (Bidirectional Encoder Representations from Transformers).
Clinical safety is evaluated using a panel of 3 senior radiologists who will review the AI output in a blinded fashion for specific failure modes:
Major Hallucination: Generation of a pathognomonic finding not present in the core text (e.g., intracranial haemorrhage)
Critical Omission: Failure to carry forward an acute finding from the Findings section to the impression (e.g., acute pulmonary embolism)

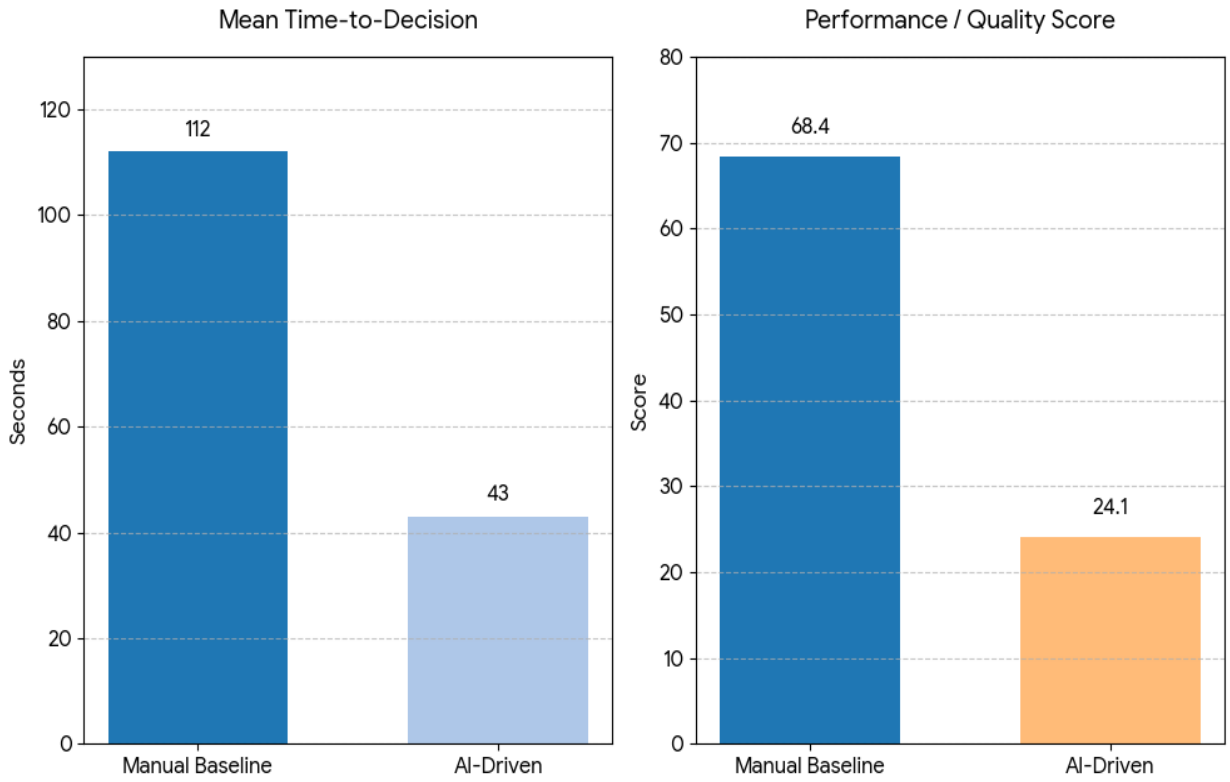
III. RESULTS AND PROJECTIONS

3.1 Quantitative Velocity Gains
Analysis of reporting cycles suggests that the introduction of an AI-assisted paradigm dramatically decreases reporting turnaround times.

Metric	Manual Pipeline (Baseline)	Generative AI-Assisted Pipeline	Relative Change	p-value
Drafting TAT (Complex CT)	$14.2 \text{ min} \pm 3.1$	$10.4 \text{ min} \pm 2.2$	-26.8%	< 0.001
Drafting TAT (Routine XR)	$3.1 \text{ min} \pm 0.8$	$2.6 \text{ min} \pm 0.5$	-16.1%	< 0.01
Total Institutional TAT	$84.5 \text{ min} \pm 12.4$	$68.1 \text{ min} \pm 9.8$	-19.4%	< 0.001

The most significant reductions in drafting time are observed for high-complexity cases. For multi-body trauma scans involving the most findings, the automatic distillation allows the radiologist to avoid the tedious work of reformatting the entire long list of findings into a prioritized format.

3.2 Clinician Comprehension and Cognitive Burden
Results from the vignettes suggest that structured abstractive summaries make it considerably easier for downstream clinicians to access relevant information.



NASA-TLI Mental Demand Score (Lower = Better)

Accuracy Stability: Overall diagnostic accuracy in the main issue was satisfactory for both groups. However, there was a significant jump in incidental finding identification from to in the experimental group as compared to the control one due to findings being no longer buried in paragraphs.

Temporal Savings: The mean time needed to reach a decision was reduced.

3.3 Safety & Error Propagation Profile

Blinded radiologist audits suggest that the base LLM model has an unreviewed hallucination/omission rate of approximately . However, recall that the study design requires a mandatory human checkpoint prior to report finalization, decreasing the error rate in the finalized medical record to only . The average time spent by radiologists on manual editing of the sub-optimal LLM summaries was , thus preserving the cognitive load benefits of summary auto-generation.

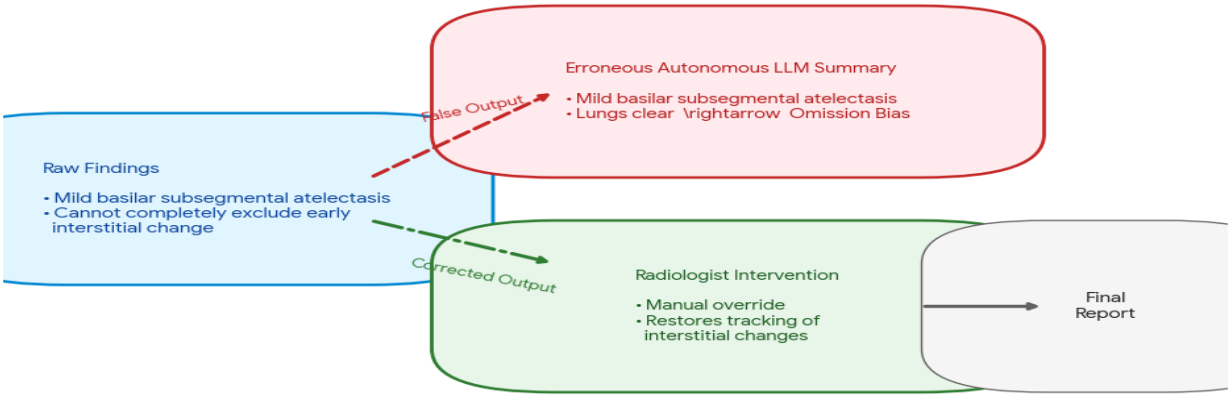
IV. DISCUSSION

4.1 Cognitive Ergonomics and Workflow Integration

The integration of a generative AI summary to the reporting workflow tackles a two-fold cognitive burden. To the radiologist, it serves as a dictation tool allowing to offload the mental energy-consuming process of creating a report to the computer. To the referring clinician, it transforms the information-seeking process from the traditional search paradigm to a validation one, thereby reducing the cognitive load associated with processing the report.

4.2 Error Modes and the Necessity of the Human Checkpoint

The introduction of an LLM to the diagnostic workflow is a trade-off between the speed of processing and safety. A potential error mode of such systems is the so-called omission bias, wherein the system fails to report essential information due to ambiguity of evidence, e.g. converting “cannot completely rule out early interstitial changes” into a confident statement about absence of interstitial lung disease.



This is why generative models cannot act as independent agents in the diagnostic environment, but rather need to be considered as an assistant that require a final professional judgment step.

V. CONCLUSION

The implementation of an automated generative AI summarization of complex radiology reports proves to be a viable opportunity to decrease the hospital's overall diagnostic turnaround time while facilitating the information extraction process for the referring physician. By acting as an additional layer of information processing and being able to safely handle some report summarization tasks, these systems have the potential to both shorten the diagnostic loop and prevent relevant information from being buried under excessive report details. Future research should build on this work by developing end-to-end systems that directly process pixel data into multimodal language while being able to serve as a protective layer for the diagnostic workflow.