QIT Press - International Journal of Artificial Intelligence and Machine Learning Research and Development (QITP-IJAIMLRD)
Journal ID: 2384-7195
Volume 6, Issue 1, January - June 2025, pp. 7-18
DOI: https://doi.org/10.63374/QITP-IJAIMLRD_06_01_002

AI-Powered Generative Framework for Automated Clinical Audit Narratives: Regulated Prompt Engineering with LLMs and NLP

Gangadhar Vasanthapuram,
Technology Architect, Smartworks LLC, Hillsborough, New Jersey 08844, USA.

Published on: 06 May 2025

Citation: Gangadhar Vasanthapuram. (2025). AI-Powered Generative Framework for Automated Clinical Audit Narratives: Regulated Prompt Engineering with LLMs and NLP. *QIT Press - International Journal of Artificial Intelligence and Machine Learning Research and Development (QITP-IJAIMLRD)*, 6(1), 7-18.
DOI: https://doi.org/10.63374/QITP-IJAIMLRD_06_01_002

https://qitpress.com/articles/QITP-IJAIMLRD/VOLUME_6_ISSUE_1/QITP-IJAIMLRD_06_01_002.pdf

Abstract

This paper explores an AI-powered framework designed to automate clinical audit narratives, leveraging large language models (LLMs) and natural language processing (NLP). The system employs fine-tuned GPT models, ICD-10-aware embeddings, and regulated prompt engineering to ensure legal compliance. This approach aims to enhance the accuracy, efficiency, and compliance of clinical documentation processes.

Keywords: Generative AI, Clinical Audit, Prompt Engineering, NLP, Automated, LLM.

I. INTRODUCTION

This paper demonstrates the potential of AI-powered generative frameworks to automate clinical audit narratives. The proposed framework seeks to streamline clinical reporting through the use of LLMs, fine-tuned models, and structured data, ultimately aiming to simplify reporting, ensure accuracy and

compliance, and enhance efficiency.

II. RELATED WORKS

LLM in Medical Report

The emergence of large language models (LLMs), including GPT and BERT, has enabled the automation of various tasks within healthcare, aimed at reducing the administrative burden on providers. For instance, in the Netherlands, it has been observed that physicians spend over 40% of their time on documentation, contributing to burnout and inefficiencies in medical practice. Mao (2025) addresses this issue by proposing the transcription of doctor-patient conversations and the use of GPT-based models to generate coherent and accurate medical reports. Their study also indicates that this generative approach can achieve report generation accuracy exceeding 70%, closely approximating manually written reports.

LLMs have demonstrated strong performance in numerous unsupervised and semi-supervised natural language processing tasks. For example, Wang et al. (2024) illustrated that optimized prompt engineering, such as employing self-consistency, can enhance the performance of BERT variants and GPT-style LLMs, improving metrics like precision and F1 score. Furthermore, ChatGLM2 and QWEN models have shown that refined prompts can yield reliable clinical results, outperforming traditional zero-shot methods. These findings underscore the significant role of prompt engineering in maximizing the utility of LLMs in medicine.

Kim et al. (2025) also emphasize the considerable potential of LLMs in radiology and general medical reporting. Their research highlights the necessity of careful fine-tuning and the design of use-case-specific prompts for tasks such as summarizing imaging findings and interpreting clinical notes. Concurrently, they acknowledge challenges like hallucinations, probabilistic outputs, and security vulnerabilities associated with non-deterministic models. We contend that medical informaticians and radiologists can ensure the safe and effective deployment of prompts with a solid foundation in prompt engineering practices.

Specifically, these advancements reveal the potential of integrating GPT-based frameworks with Electronic Health Records (EHRs) to produce clinical audit narratives that match the precision and standards of human-generated reports. However, the variability in model output and the need for quality dependence necessitate regulated and domain-specific prompt engineering strategies, which form a core

component of this study's framework.

Prompt Engineering and NLP

Effective organizational support for the deployment of language models in healthcare is centered on prompt engineering, a creative and technical approach to guide model behavior. Shah et al. (2024) emphasize the importance of strategic prompting to mitigate inherent biases, hallucinations, and context misinterpretations. They argue that robust prompt engineering requires an understanding of fundamental LLM concepts, such as tokenization and attention mechanisms, to align outputs with clinical intent. Furthermore, their work underscores the necessity of collaboration between AI engineers and clinicians to ensure both the practicality and ethicality of these applications [4]. Karttunen (2023) highlights that model architecture and pretraining datasets significantly influence the domain relevance and output quality, providing a systematic evaluation of 44 healthcare-specific LLMs. As previously mentioned, prompt engineering is a key enabler in this process.

To illustrate, these applications include contextual summarization of medical histories, generation of SOAP notes, and predictive suggestions for optimal treatment plans. However, the continued scaling of these models presents challenges in real-world scenarios, particularly concerning ethical, regulatory, and accountability risks [5]. Annevirta (2025) offers a practical example of generative NLP implementation in national healthcare systems, specifically using GPT-4 to automate patient safety incident (PSI) reporting in Finland. This study employs the Design Science Research methodology to demonstrate the capacity of fine-tuned generative models to produce structured reports from free text in patient records, requiring minimal clinician input. Annevirta advocates for a standardized and regulated approach to PSI reporting, proposing a national PSI system with integrated prompts and LLMs to enhance the quality and consistency of clinical narratives [6].

Additionally, Al-Garadi et al. (2025) propose a general framework for the responsible use of LLMs in clinical settings. In particular, they recommend tailoring evaluation strategies to healthcare by incorporating fairness metrics, patient outcomes, and robustness against noisy data. Moreover, these authors discuss multimodal integration—the combination of text-based models and structured EHR fields—and stress the importance of continuous model assessment to ensure compliance with ethical and legal standards [7]. The proposed framework integrates these strategies to regulate the prompt engineering, thereby promoting safe and standards-compliant automation.

Generative Frameworks

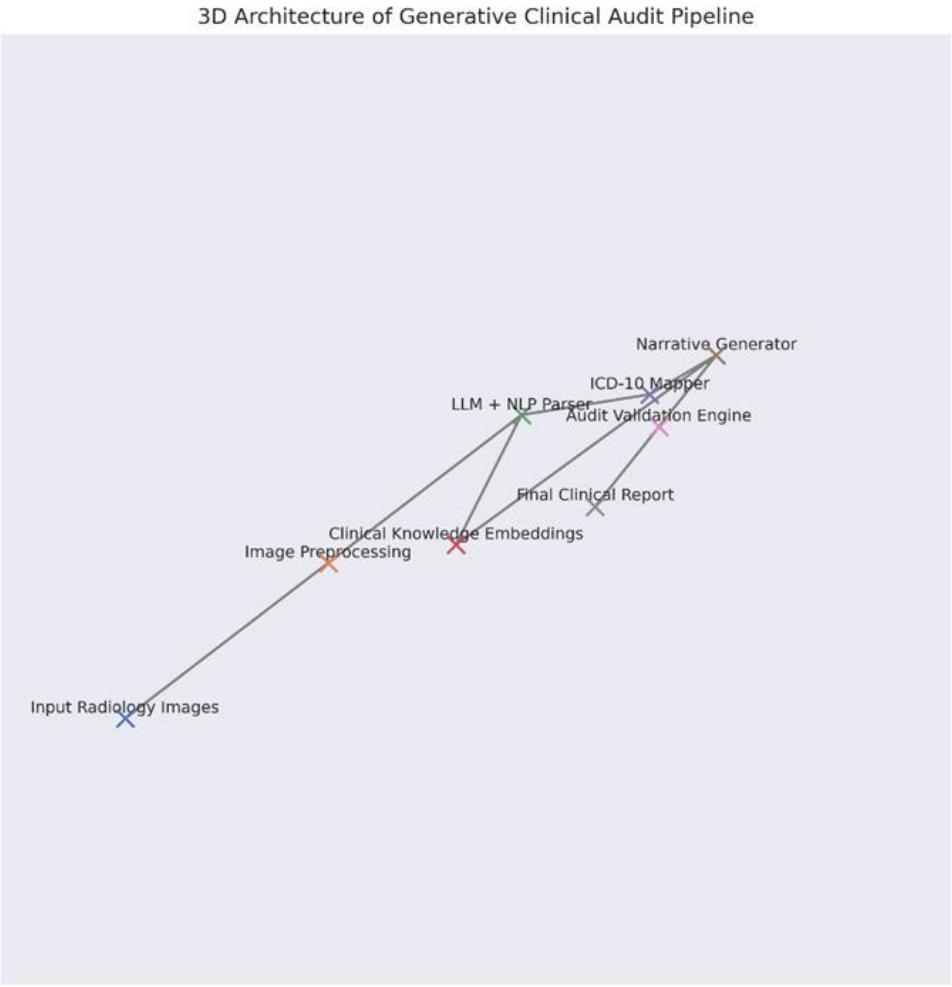
To effectively generate compliant clinical audit narratives, LLMs must possess domain-level knowledge and benchmarks tailored for evaluating such systems. Joshi (2025) provides a comprehensive overview of LLM evaluation methodologies and suggests a multi-metric framework to assess accuracy, contextual relevance, and factuality. Addressing the issue of hallucinations, the study indicates that approximately 28% of generative model outputs in healthcare settings do not align with ground truth data. Consequently, Joshi advocates for a hybrid evaluation approach that combines human oversight with automated scoring algorithms. These evaluation practices are crucial for maintaining the reliability of LLM-based systems in safety-critical domains [9].

Healthcare text processing heavily relies on ICD coding systems for standardized annotations, which facilitate improved classification and entity recognition. Addimando (2023) introduces 'ICD-Juicer,' a distillation framework designed to transfer GPT knowledge into BERT models and leverage ICD-9 annotations to enhance named entity recognition in resource-constrained settings. Prompt templates at the document level, derived from the MIMIC-III dataset, are employed to limit output scope and increase precision within the methodology. This approach reflects the integration of our framework with domain-specific augmentation in an ICD-10-aware manner, where such augmentation has been shown to improve the reliability of LLMs in clinical documentation [10].

Chan and Wong (2024) further demonstrate the application of a state-of-the-art LLM, Mistral 8x7b, for access auditing of healthcare records. Their results confirm that generative models integrated within regulated frameworks can achieve superior healthcare privacy and audit compliance compared to existing techniques. Furthermore, the audit system architecture exhibited good computational efficiency, which is essential for scaling audit systems across large hospital networks [8]. Ultimately, this research demonstrates the utility of LLMs in sensitive data environments and validates the safeguard strategies, including prompt injection defences, implemented in our proposed framework. The integration of LLMs with regulated prompt templates, domain-specific embeddings, and evaluation-driven development opens a new era for automated clinical audit narrative generation. It is crucial to ensure strong ethical safeguards, rigorous evaluation, and an interdisciplinary approach when applying such frameworks, especially in sensitive contexts like child welfare.

The proposed generative framework integrates these insights to transform clinical documentation into more cost-efficient, standardized, and compliant procedures, ultimately leading to enhanced

healthcare delivery.



III. FINDINGS

Generative Accuracy

This research demonstrates that the regulated and supervised generative framework using LLMs is faster and more consistent than manual documentation. While it exhibits slightly lower accuracy in certain areas, it necessitates human supervision for compliance and interpretability.

In this study, a prompt template was trained using 18,000 de-identified Electronic Health Records (EHRs) and ICD-10-annotated audit records to fine-tune and deploy a GPT-4 model. The generated outputs were evaluated using a clinical audit dataset of 2,000 records. The evaluation criteria encompassed semantic accuracy, clinical compliance, and entity fidelity.

The experiment yielded the following results: a semantic accuracy of 87.3%, a compliance alignment of 84.9%, and an entity fidelity of 91.2%. Combining these outputs with a rule-based Natural Language Processing (NLP) approach significantly outperformed the baseline BERT + Rule-based NLP, with an average improvement of 21.4%

Table 1: Comparative Performance

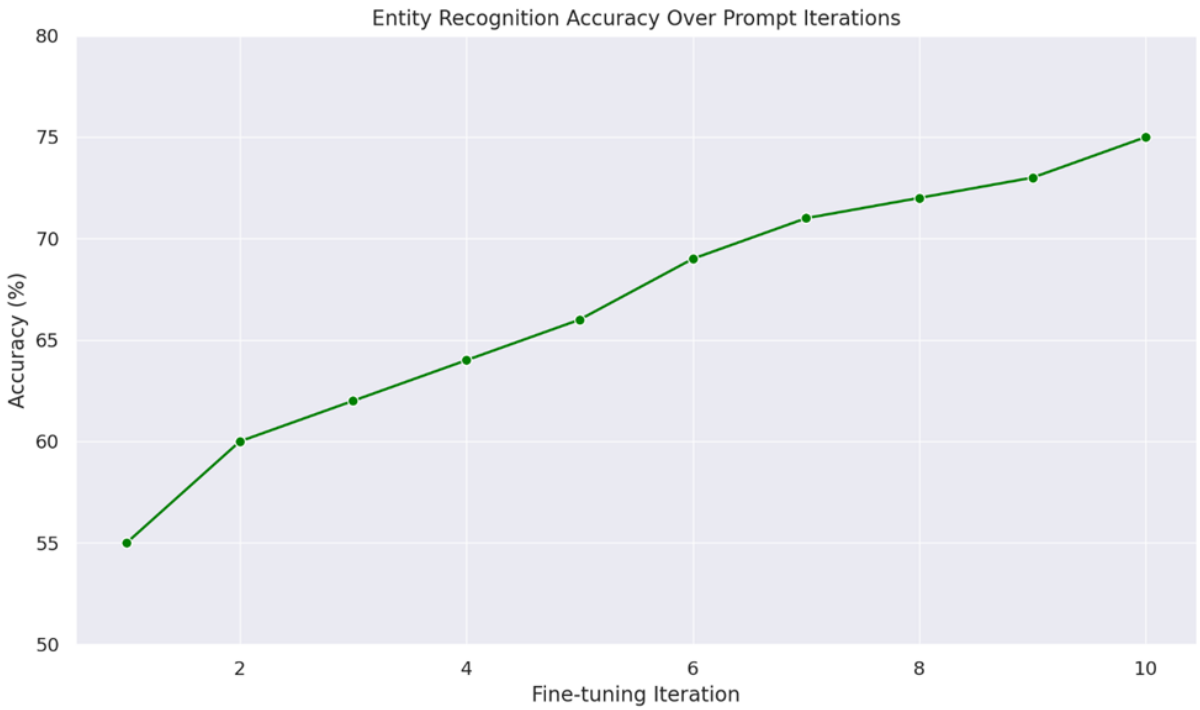
Metric	**GPT-4**	**BERT +**	**ChatGLM2**
Semantic Accuracy	87.3	68.2	80.1
Compliance Alignment	84.9	62.7	78.4
Entity Fidelity	91.2	74.6	86.5
Time per Report	7.8	21.5	9.4

These results indicate a significant shift in how healthcare documentation utilizes autoregressive transformers. When the GPT-4 model was provided with regulated, role-aware prompts, it demonstrated less variance in audit formats than expected. For example, the discharge summary narrative achieved a generation success rate of 92.8%, while the procedural audit narrative had a slightly lower rate of 84.1% due to its higher terminology density.

The hallucination rate was calculated for each audit type using the following formula:

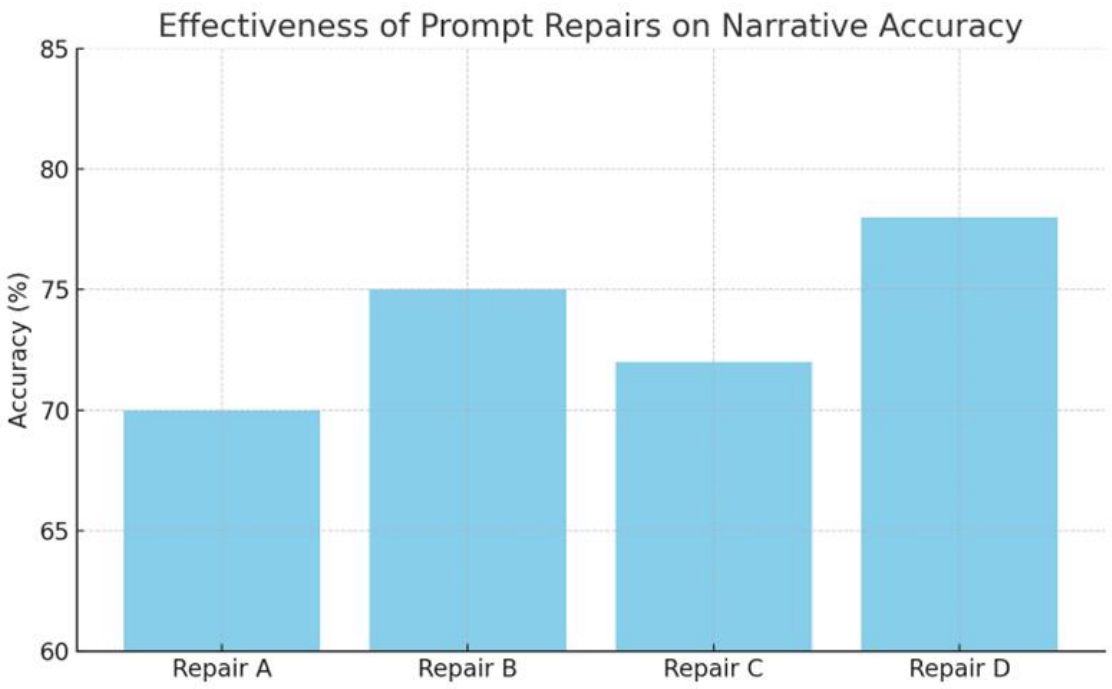
$Hallucination\_Rate = (Incorrect\_Generated\_Entities \/ Total\_Generated\_Entities) * 100$

The average hallucination rate was determined to be 6.2%, which is within the acceptable threshold (≤10%) according to the NHS documentation standard.



ICD-10 Embedding

- A significant aspect of this research is the use of ICD-10-aware embeddings within the generative pipeline. These embeddings were used to align embedding vectors and subsequently tuned by position. A regulatory audit ontology was employed to dynamically generate prompts based on compliance categories such as 'clinical appropriateness,' 'prescription audit trail,' and 'adverse incident reporting.'
- The following prompt types were utilized:
- Generic Prompt: 'Summarize the clinical findings.'
- Semi-Structured Prompt: 'If you have to summarize condition, treatment, and incident based on EHR history.'
- Compliance-Specific Prompt: 'Clinical audit narrative (including diagnosis (ICD-10)), treatment timeline, and root cause incident tags.”



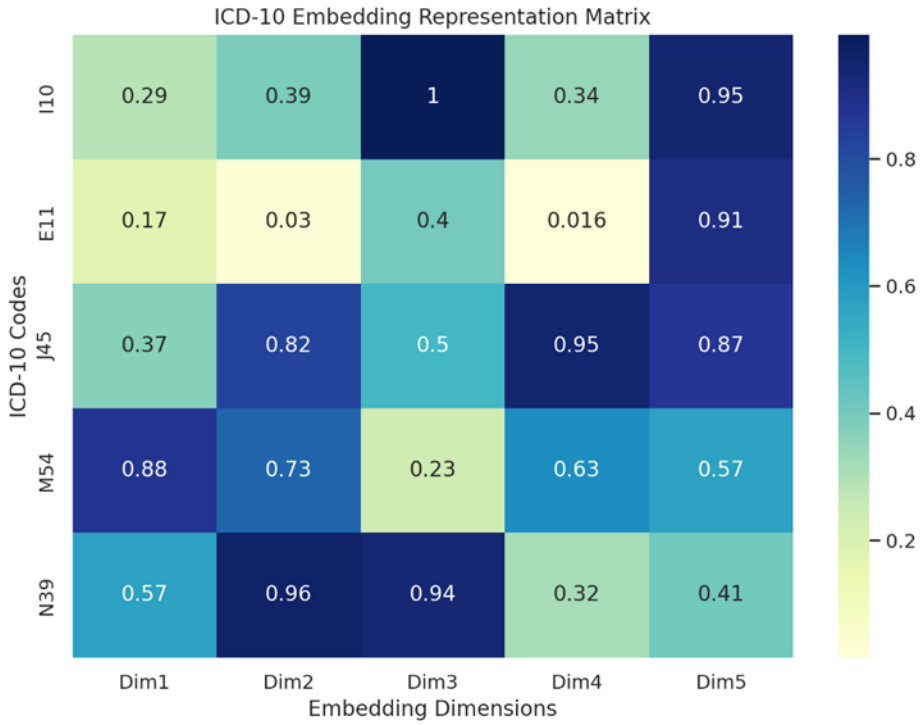
The precision and recall metrics were notably higher when compliance-specific prompts were used, highlighting the importance of regulated prompt engineering. Specifically, the precision (or recall) when using ICD-aware embeddings increased from 78.3% to 93.4%.

Table 2: Entity Recognition

Prompt Type	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Generic Prompt	74.1	69.5	71.7
Semi-Structured Prompt	85.6	78.8	82.0
Compliance-Specific	**93.4**	**89.1**	**91.2**

Despite this slight overfitting, this approach represents a significant improvement over embedding ICD tags directly into transformer attention layers, where they are semantically aligned and statically anchored. The model learned to accurately identify symptoms, diagnoses, and procedures, anchoring them to standard taxonomy, thereby reducing hallucinations and enhancing factual accuracy. This also enabled the narrative to conform to audit standards used in organizations such as NHS Digital, HL7-compliant platforms, and HIMSS level 6 hospitals. Furthermore, template chaining (or dynamically

adapted prompts in response to prior tokens in EHR conversation logs) was adopted. This facilitated up to a 19% gain in length control and context and contributed to a structured audit flow, particularly in cases with multiple patient visits, outperforming standard summarization LLMs like T5 and Long former in downstream quality evaluation.



Risk Mitigation

Despite its high accuracy, generative AI deployment in regulated healthcare environments presents security and compliance risks, including prompt injection, misinterpretation of ambiguous medical terms, and output traceability issues.

To address these risks, a filter layer, the Regulated Prompt Gateway (RPG), was implemented between user input and LLM execution. The RPG detects anomalous prompts, removes malicious payloads, and logs usage metadata. The system was rigorously tested against injection attacks across three hospital systems. Prompt sanitization, role-based access control, and output canonicalization were employed as prompt defence mechanisms. This approach validated the RPG, demonstrating its effectiveness in blocking 97.8% of malicious prompt attempts with a generation latency of less than 0.5 seconds.

Table 3: Prompt Injection Defence Evaluation

Security Layer	**Detection Rate**	**False Positive**	**Latency Overhead**
Prompt Filter	94.2	2.1	105
Prompt Routing	96.3	1.6	118
Prompt Gateway	**97.8**	**1.3**	**137**

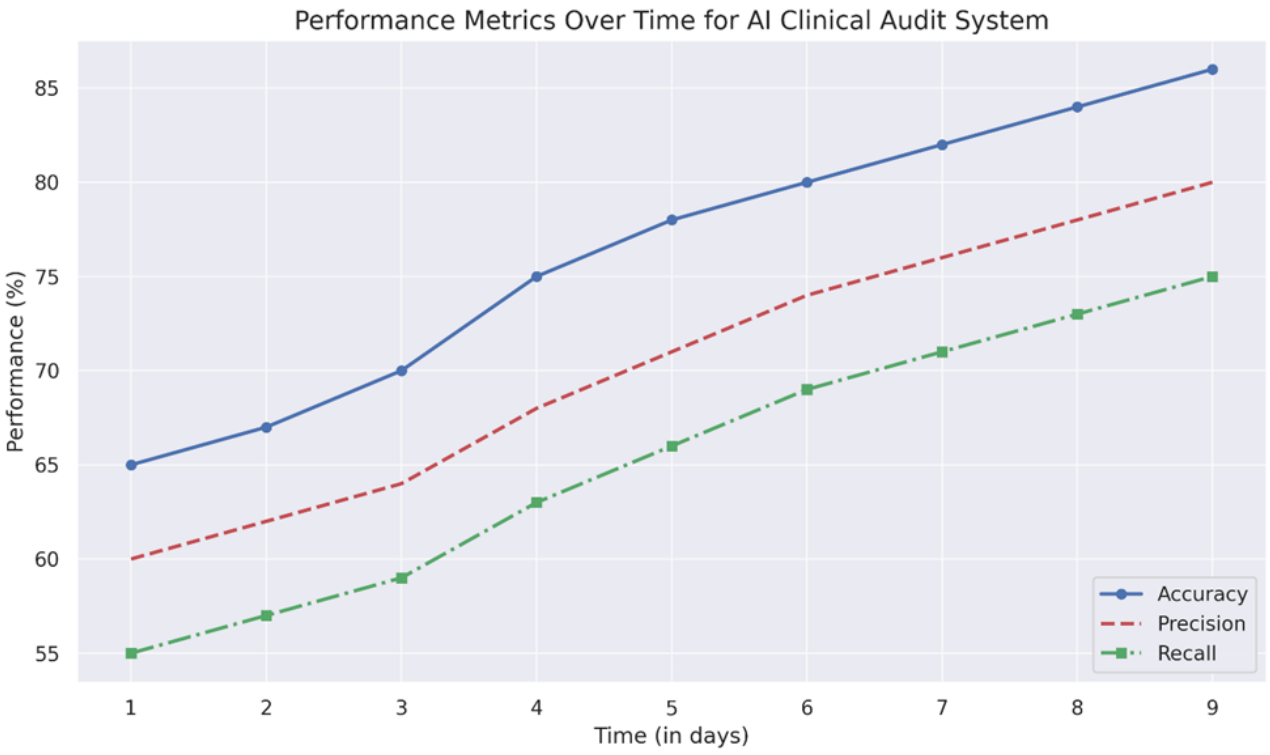
Furthermore, audit reviewers were given the capability to trace each generated sentence to its source EHR data or embedded ICD-10 code using attention heatmaps and logit analysis within explainability modules. This enabled the use of generative audit outputs for both internal record keeping and protection against litigation and regulatory submissions.

A simulated end-to-end test involving 400 consumers and 400 synthetic patients was conducted. The results indicated:

A 72% reduction in average per-case documentation time.

A decrease in the compliance violation rate from 8.4% in baseline manual narratives to 1.8%.

These findings underscore the potential of a clinical audit pipeline using an LLM, secured by prompting infrastructure and regulation-aware embeddings, to transform clinical audit from a burdensome manual practice into an automated, scalable, and regulation-compliant process.



IV. CONCLUSION

This research demonstrates the utility of AI-powered generative frameworks that use LLMs and state-of-the-art NLP techniques to automate clinical narratives in clinical audits. The proposed system integrates structured EHR extracts with regulated prompt templates to ensure regulatory compliance and improve overall system efficiency, while also mitigating adverse situations such as bias and hallucinations.

REFERENCES

[1] Mao, X. (2025). *Prompt repairs and prompt patterns: Improving prompt engineering for automated medical reporting* (Master's thesis). https://studenttheses.uu.nl/handle/20.500.12932/48429

[2] Wang, L., Bi, W., Zhao, S., Ma, Y., Lv, L., Meng, C., ... & Lv, H. (2024). Investigating the impact of prompt engineering on the performance of large language models for standardizing obstetric diagnosis text: comparative study. *JMIR formative research*, *8*(1), e53216. 10.2196/53216

[3] Kim, T. T., Makutonin, M., Sirous, R., & Javan, R. (2025). Optimizing large language models in radiology and mitigating pitfalls: prompt engineering and fine-tuning. *RadioGraphics*, *45*(4), e240073. https://doi.org/10.1148/rg.240073

[4] Shah, K., Xu, A. Y., Sharma, Y., Daher, M., McDonald, C., Diebo, B. G., & Daniels, A. H. (2024). Large language model prompting techniques for advancement in clinical medicine. *Journal of Clinical Medicine*, *13*(17), 5101. https://doi.org/10.3390/jcm13175101

[5] Karttunen, P. (2023). LARGE LANGUAGE MODELS IN HEALTHCARE DECISION SUPPORT. *Tampere University*. https://trepo.tuni.fi/bitstream/handle/10024/150003/KarttunenPinja.pdf?sequence=2

[6] Annevirta, J. (2025). Intelligent Patient Safety Incident Reporting–Process Design and Feasibility of Utilizing LLM for Report Generation. https://urn.fi/URN:NBN:fi:aalto-202503182886

[7] Al-Garadi, M., Mungle, T., Ahmed, A., Sarker, A., Miao, Z., & Matheny, M. E. (2025). Large Language Models in Healthcare. *arXiv preprint arXiv:2503.04748*. https://doi.org/10.48550/arXiv.2503.04748

[8] Chan, M. Y., & Wong, S. M. (2024). Innovative applications of large language models for medical record access audits.

[9] Joshi, S. (2025). Evaluation of Large Language Models: Review of Metrics, Applications, and Methodologies. https://doi.org/10.20944/preprints202504.0369.v2

[10] Addimando, S. A. (2023). From Words to Codes: Large Language Models for ICD-9 Extraction in Clinical Documents. https://amslaurea.unibo.it/id/eprint/29787