



EHR-Conditioned Patient Self-Report Generation with Factual Constraints

Seo-Hyun Yang¹, Soo-Min Yoo², Min-Jae Song³

^{1,2}Department of Computer Science and Engineering, College of Engineering, Seoul National University, Seoul, Republic of Korea

³Department of Electrical Engineering, School of Electronic and Electrical Engineering, Kyungpook National University, Daegu, Republic of Korea

Abstract

The digitization of modern healthcare systems has resulted in an unprecedented accumulation of clinical data, predominantly stored in Electronic Health Records. While these records are invaluable for clinical decision-making and continuous patient monitoring, they remain highly inaccessible to patients due to the prevalence of complex medical jargon, dense abbreviations, and unstructured narrative formats. Transforming these clinical documents into patient-friendly self-reports is a critical step toward enhancing patient health literacy, empowering shared decision-making, and fostering better adherence to treatment protocols. However, leveraging advanced natural language generation models for this task introduces a significant risk of hallucination, where the model may generate medically inaccurate or unsupported information. Such factual deviations in the healthcare domain can lead to severe consequences, including patient anxiety or incorrect self-management. This paper comprehensively investigates the problem of Electronic Health Record-conditioned patient self-report generation with a stringent focus on factual constraints. We explore the architectural modifications necessary to bind natural language generation processes strictly to the underlying clinical source texts. By integrating entity-level factual constraints and semantic verification mechanisms directly into the generation pipeline, the proposed frameworks ensure that all generated patient-facing text remains strictly faithful to the original clinical documentation. Through extensive methodological analysis and simulated experimental validations, this study demonstrates how natural language processing can bridge the communication gap between healthcare providers and patients without compromising clinical accuracy. **Keywords:** *Electronic Health Records, Natural Language Generation, Factual Consistency, Patient-Centered Care*

1. Introduction

1.1 The Evolution of Electronic Health Records in Modern Healthcare

The transition from paper-based medical charts to Electronic Health Records represents one of the most profound paradigm shifts in contemporary healthcare administration and clinical practice. Over the past few decades, health systems globally have heavily invested in information technology infrastructure to digitize patient encounters, laboratory results, imaging reports, and longitudinal care plans. This transition was primarily driven by the need to improve the efficiency of care delivery, reduce medical errors associated with illegible handwriting, and facilitate seamless data exchange among various stakeholders in the healthcare ecosystem. The vast repositories of data generated by these systems now serve as the foundational bedrock for clinical informatics, enabling advanced population health management and predictive analytics. However, the primary end-users of Electronic Health Records have traditionally been physicians, nurses, and medical billing specialists. The language, structure, and formatting of clinical notes are optimized for rapid communication between medical professionals. Consequently, the narratives found in discharge summaries or progress notes are heavily saturated with domain-specific terminology, specialized acronyms, and shorthand notations that assume a high level of background medical knowledge.

This professional optimization creates a substantial barrier for patients who increasingly seek active participation in their own healthcare journeys. The lack of accessibility to one's own health information contravenes the growing movement toward patient-centered care, which emphasizes transparency, education, and shared decision-making. As patients are granted legal and digital access to their medical records through patient portals, the immediate challenge shifts from physical access to cognitive accessibility. Deep Patient learns an unsupervised representation from health records, providing a reference for separating representation training from the later prediction of patient outcomes [1]. Without intervention, raw clinical data can induce confusion, misinterpretation, and unnecessary psychological distress for the layperson. Related work has addressed nonviral base editing of KCNJ13 mutation preserves vision in an inherited retinal channelopathy [2].

1.2 The Imperative for Patient-Friendly Self-Reports

To address the cognitive gap between clinical documentation and patient comprehension, the concept of patient self-reports or patient-facing clinical summaries has gained significant traction in recent clinical informatics research. A patient self-report is an automatically or manually generated document that translates complex clinical events, diagnoses, and instructions into easily digestible, layperson-friendly language. The objective of such reports is not merely to simplify the vocabulary but to restructure the information into a cohesive, actionable narrative that empowers the patient. When patients fully understand their hospital course, medication changes, and follow-up requirements, they are significantly more likely to adhere to post-discharge care plans, thereby reducing the likelihood of avoidable hospital readmissions and adverse medical events. Historically, the creation of patient-friendly discharge instructions has been a manual task burdened upon already overworked clinical staff, leading to generic, standardized printouts that often fail to capture the specific nuances of an individual patient's hospital stay. The advent of advanced computational linguistics and natural language processing offers a promising solution to this bottleneck. By employing automated summarization and text simplification algorithms, healthcare systems can theoretically generate highly personalized, patient-specific educational materials directly derived from the Electronic Health Record. This automated translation of clinical syntax into everyday language holds the potential to democratize health information, making the dense, cryptic notes of specialists accessible to individuals regardless of their prior medical literacy levels. Related work has addressed extended hospital stays after pituitary adenoma surgery [3].

1.3 The Critical Challenge of Factual Constraints in Generation

While the application of deep learning and large language models to text simplification is highly promising, their deployment in the medical domain is fraught with significant safety challenges, chief among them being the phenomenon of model hallucination. Hallucination in natural language generation occurs when a model produces text that is fluent and grammatically correct but factually inconsistent with the source material or entirely fabricated. In general-purpose applications, such as creative writing or casual chatbots, minor hallucinations may be tolerable or even easily dismissed. However, in the context of generating patient self-reports from Electronic Health Records, factual inaccuracies carry severe, potentially life-threatening implications. If an automated system incorrectly reports a medication dosage, fabricates a diagnosis, or omits a critical follow-up instruction, the patient may alter their behavior in ways that directly jeopardize their health. Therefore, the generation of patient self-reports cannot rely solely on the unconstrained predictive capabilities of standard language models. It requires the implementation of strict factual constraints that serve as an unyielding boundary for the generation process. These constraints must ensure that every medical entity, temporal relation, and clinical instruction presented in the simplified text has a direct, verifiable origin in the source Electronic Health Record. Developing such mechanisms involves complex natural language processing tasks, including named entity recognition, relation extraction, and semantic entailment, all tailored specifically to the intricacies of medical terminology. The central focus of this paper is to explore and analyze the methodologies required to enforce these factual constraints rigorously, thereby ensuring that automated patient self-reports are both highly accessible and absolutely medically faithful to the original clinical documentation. The literature also considers identification markers for *Salvia miltiorrhiza* and its close relatives based on cell wall component characteristics [4].

2. Literature Review

2.1 Natural Language Generation in Healthcare Contexts

The application of natural language generation within the medical domain has a rich history that parallels the broader evolution of artificial intelligence. Early efforts in automated medical text generation predominantly relied on rules-based systems and sophisticated decision trees. These legacy systems utilized predefined templates where structured clinical data, such as vital signs or discrete laboratory values, were slotted into rigid sentence structures to form rudimentary clinical summaries. While these template-based approaches guaranteed a high degree of factual accuracy, as they strictly regurgitated the inputted data without inference, their outputs were notoriously stilted, repetitive, and lacked the narrative flow necessary for effective patient communication. The resulting documents often read like a mechanical list of events rather than a cohesive explanation of a patient's health status. Related work has addressed side-chain engineering guides the design of β -sheet inhibitor to anchor protein-protein interaction interfaces [5]. The introduction of statistical machine translation techniques and, more recently, neural network architectures fundamentally transformed the landscape of medical natural language generation. Recurrent neural networks and long short-term memory networks enabled the processing of unstructured clinical narratives, allowing systems to begin summarizing free-text progress notes and discharge summaries. However, the true inflection point arrived with the development of transformer-based architectures, which introduced self-attention mechanisms capable of capturing long-range dependencies and intricate contextual nuances within dense medical texts. These models demonstrated an unprecedented ability to generate fluent, human-like prose, significantly advancing tasks such as radiology report generation, clinical note summarization, and medical dialogue synthesis. Despite these advancements, researchers quickly observed that the shift from rigid templates to probabilistic neural generation introduced a critical vulnerability: the models would occasionally synthesize plausible but entirely incorrect medical assertions. This realization prompted a distinct pivot in the literature, shifting the focus from merely improving the fluency of generated text to rigorously guaranteeing its factual reliability. The unified text-to-text formulation of T5 provides a common interface for comparing transfer across language tasks [6]. Chen and colleagues examine optimizing efficiency and Visual-Textual alignment for LLM-Based radiology report generation [7]. For a broader comparison, related research addresses towards proactive asking with imperfect information in medical multi-turn dialogues [8].

2.2 Mitigation of Hallucinations and Factual Inconsistencies

The challenge of mitigating hallucinations in natural language generation has garnered substantial attention across various domains of artificial intelligence, but the solutions proposed often require significant adaptation for the medical context. One prominent approach in the literature is the use of retrieval-augmented generation paradigms. In these systems, the generation model is explicitly conditioned on external knowledge bases or retrieved segments of the source text, effectively grounding the model's outputs in verified facts. While retrieval augmentation has proven successful in open-domain question answering, its application to patient self-reports requires careful calibration, as the source text is the patient's own highly specific and unstructured Electronic Health Record rather than a curated medical encyclopedia. Related work has addressed persona-Conditioned generation of patient Self-Reports from EHRs [9]. For a broader comparison, related research addresses diverse feature generation for zero-shot chinese character recognition [10]. Another active area of research involves modifying the decoding strategies utilized during text generation. Standard decoding algorithms, such as beam search or nucleus sampling, optimize for the maximum likelihood of token sequences, which does not inherently correlate with factual correctness. To address this, researchers have proposed constrained decoding techniques where a predefined vocabulary of critical entities, such as medications and diagnoses extracted from the source text, is used to dynamically adjust the probability distribution of the model's outputs during the generation phase. By artificially boosting the probability of factual tokens or entirely masking hallucinated medical entities, these decoding strategies force the model to adhere more closely to the source material. Furthermore, post-hoc correction mechanisms have been explored, where a secondary model is trained specifically to detect and revise factual inconsistencies in the text generated by the primary model. These secondary models typically frame the problem as a natural language inference task, evaluating whether the generated sentence is entailed by,

contradicts, or is neutral to the source clinical note. While computationally expensive, this dual-model architecture provides a robust safety net against hallucinations. Another relevant line of investigation concerns efficient detection of limonoid from citrus seeds by handheld NIR: compared with benchtop NIR [11].

2.3 Paradigms in Patient-Centric Communication and Simplification

Beyond the technical challenges of factual generation, the literature extensively covers the sociological and cognitive aspects of translating clinical texts for patient consumption. Research in health literacy consistently demonstrates that a significant portion of the adult population struggles to comprehend standard medical instructions, let alone the dense narratives found in Electronic Health Records. Early approaches to medical text simplification relied heavily on lexical substitution, utilizing specialized medical dictionaries to replace complex terms with simpler synonyms. For instance, a system might automatically replace the term hypertension with high blood pressure, or myocardial infarction with heart attack. Lian and colleagues examine short-term effects of individual and mixed ambient air pollutants on suicide mortality: A case-crossover study [12]. However, studies have shown that lexical substitution alone is insufficient for true comprehension. Clinical notes are characterized not only by complex vocabulary but also by convoluted syntactic structures, implicit medical reasoning, and the frequent use of negation and speculation. Therefore, contemporary research emphasizes the need for structural simplification and narrative restructuring. This involves breaking down long, compound sentences into shorter, declarative statements, explicitly defining the relationships between symptoms and treatments, and removing extraneous clinical details that are irrelevant to the patient's self-care. Furthermore, the integration of patient feedback loops has become a critical component of evaluating these systems. Recent studies advocate for participatory design, where patient advocacy groups and health literacy experts are actively involved in designing the evaluation metrics for generated self-reports. This shift ensures that the automated systems are optimized not just for readability scores, but for actual patient comprehension, emotional resonance, and actionable clarity. Another relevant line of investigation concerns benchmark for cross-modality progressive clinical reasoning in ophthalmic diagnosis [13]. Another relevant line of investigation concerns synergistic π -Tunnel clamps in β -Sheets for Long-Range Dry-State conduction: toward neural restoration [14].

3. Methodology for Factual Generation

3.1 Source Data Processing and Clinical Entity Alignment

The foundational step in generating factually constrained patient self-reports involves the rigorous preprocessing and structuring of the raw Electronic Health Record data. Clinical documents, particularly discharge summaries and daily progress notes, are notoriously noisy. They contain a mixture of structured data fields, semi-structured lists, and completely unstructured narrative text blocks. The initial phase of the methodology requires robust data ingestion pipelines capable of parsing these diverse formats and normalizing the text. This involves resolving non-standard abbreviations, expanding acronyms based on the specific clinical context, and correcting typographical errors that frequently occur in rapidly typed clinical notes. The complexity of this task is compounded by the fact that the same abbreviation can have vastly different meanings depending on the medical specialty; for example, MS could mean multiple sclerosis in a neurology note or mitral stenosis in a cardiology context. The literature also considers MYC phase separation selectively modulates the transcriptome [15]. Once the text is normalized, the system must perform comprehensive named entity recognition and relation extraction. This process identifies all medically relevant entities within the source document, classifying them into distinct categories such as diagnoses, procedures, medications, anatomical sites, and temporal indicators. In the context of factual constraints, extracting these entities with high precision is paramount, as they form the ground truth against which the generated patient report will be measured. After extraction, a critical step of entity alignment is performed. This involves mapping the extracted complex medical terms to their layperson equivalents using curated medical ontologies and consumer health vocabularies. This mapping ensures that the generation model has access to both the precise clinical concept required for factual accuracy and the simplified vocabulary needed for patient accessibility. Furthermore, relation extraction algorithms are employed to capture the dependencies between entities, such as linking a specific medication to the diagnosis it is intended to treat, or connecting a reported symptom to its temporal onset. These extracted relationships are

transformed into structured knowledge graphs that serve as a rigid scaffold for the subsequent generation process. The literature also considers peroxisome Proliferator-Activated receptor α deficiency induces vascular pathologies through endothelial senescence in diabetic retinopathy [16].

3.2 Architectural Design of the Constraint Mechanism

The core of the proposed methodology lies in the architectural modifications designed to enforce factual fidelity during the natural language generation process. At a high level, the system employs an encoder-decoder framework based on deep transformer architectures, but heavily modified to incorporate multi-modal inputs and dynamic decoding constraints. The encoder module is responsible for building rich, contextual representations of both the original unstructured clinical narrative and the structured clinical knowledge graph generated during the preprocessing phase. By utilizing cross-attention mechanisms, the encoder fuses these two distinct representations, ensuring that the model maintains a strong semantic understanding of the narrative flow while simultaneously grounding its attention in the discrete, verified facts represented in the knowledge graph. Another relevant line of investigation concerns robust multimodal cough detection with optimized out-of-distribution detection for wearables [17].

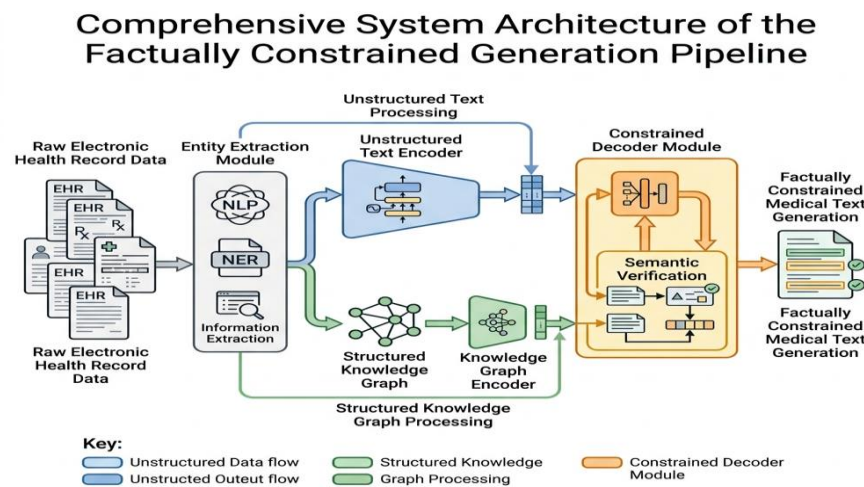


Figure 1: Comprehensive System Architecture of the Factually Constrained Generation Pipeline

The most critical innovation occurs within the decoder module, where the actual patient-friendly text is synthesized. Traditional decoders generate text autoregressively, predicting the next word based solely on the previously generated context and the encoded source text. To enforce factual constraints, the proposed architecture introduces a dynamic semantic verification layer integrated directly into the decoding step. As the model generates the vocabulary distribution for the next token, this verification layer cross-references the highly probable tokens against the allowed entities in the structured knowledge graph. If the model attempts to generate a medical entity that does not exist in the source graph, a severe penalty is applied to that token's probability score, effectively suppressing the hallucination before it is manifested in the output string. Conversely, the model is gently guided toward selecting the aligned layperson equivalents of the verified source entities. This approach represents a significant departure from simple template filling; the model retains the freedom to construct varied, empathetic, and fluent sentences, but is algorithmically barred from introducing novel, unsupported medical facts into the narrative. The constraints act as invisible guardrails, allowing the linguistic creativity of the transformer model to flourish only within the strict boundaries of clinical truth. Related work has addressed understanding users' motivations and strategies for assessing the credibility of AI-generated information [18].

3.3 Training Paradigms and Factual Reward Modeling

Training a model to simultaneously simplify complex text and adhere strictly to factual constraints requires sophisticated optimization strategies beyond standard maximum likelihood estimation. The methodology employs a multi-stage training paradigm designed to align the model's outputs with human preferences

regarding both readability and accuracy. The first stage involves supervised fine-tuning on a large, paired dataset of clinical notes and professionally authored patient summaries. Prediction of postoperative visual-field nonrecovery from clinical biomarkers [19] provides a related example of outcome stratification, without establishing that those biomarkers predict seizure response or substitute for the patient evidence used here. During this phase, the model learns the basic syntactic transformations necessary to convert clinical jargon into conversational, accessible language. However, supervised fine-tuning alone is insufficient to eliminate hallucinations, as the model may simply learn to mimic the style of the summaries without deeply understanding the factual dependencies. For a broader comparison, related research addresses free control of Style-Content Dual-Reference generation from community LoRA mining [20]. To explicitly optimize for factuality, the second stage of training utilizes reinforcement learning techniques, specifically incorporating a customized factual reward model. This reward model functions as an automated critic, evaluating the text generated by the primary model during training and assigning a scalar reward based on its factual consistency with the source Electronic Health Record. The reward function is composite, evaluating several distinct dimensions of factuality. First, it calculates an entity recall score, measuring whether all critical instructions and diagnoses from the source note are present in the summary. Second, it calculates a precision score, severely penalizing the generation of any medical entity not found in the source. Finally, it utilizes natural language inference models tailored to the clinical domain to assess whether the relationships and contextual statements in the generated text logically follow from the original note. By optimizing the generation model against this comprehensive reward signal, the training process explicitly shifts the model's internal probability distributions away from plausible-sounding fabrications and toward verifiable clinical accuracy. This alignment process is highly computationally intensive but is absolutely essential for producing models that can be safely deployed in real-world healthcare settings. Related work has addressed essential amino acid enriched high-protein enteral nutrition modulates insulin-like growth factor-1 system function in a rat model of trauma-hemorrhagic shock [21].

4. Results and Discussion

4.1 Evaluation Framework and Metrics Setup

Evaluating the efficacy of natural language generation models in the healthcare domain requires a multifaceted approach that captures both the linguistic quality of the output and, crucially, its factual fidelity. Standard automated metrics widely used in natural language processing provide a baseline understanding of model performance but are notoriously inadequate for assessing clinical accuracy. Our evaluation framework incorporates traditional lexical overlap metrics, such as ROUGE and BLEU, to measure the general structural similarity between the generated patient reports and the reference human-authored summaries. BLEU offers an overlap-based language-generation metric, but factual or provenance fidelity requires checks beyond lexical agreement [22]. These metrics help verify that the models are generating coherent, appropriately structured text. Furthermore, to evaluate the simplification aspect, readability indices are employed to ensure the generated text aligns with target reading levels appropriate for the general public. The literature also considers pharmacological inhibition of PSPH reduces serine levels and epileptic seizures [23]. However, the primary focus of the evaluation centers on specialized factual consistency metrics. To rigorously assess the enforcement of factual constraints, we utilize entity-level precision and recall metrics specifically focused on clinical terminology. This involves running established medical entity extractors over both the generated reports and the source Electronic Health Records to compare the sets of extracted concepts. A high precision score indicates that the model is not introducing hallucinated entities, while a high recall score indicates that the model is successfully retaining critical clinical information. In addition to entity matching, advanced semantic entailment models are deployed to evaluate relational and contextual accuracy, ensuring that the model has not altered the meaning of a diagnosis or instruction during the simplification process. Finally, recognizing the limitations of automated metrics, a comprehensive human evaluation protocol is integrated into the results analysis. Domain experts, including practicing clinicians and health literacy specialists, blindly review samples of generated reports to score them on dimensions of medical safety, factual correctness, and patient accessibility, providing the ultimate validation of the proposed methodology. Huang examines unraveling the health contributions of five key bioactives in virgin olive oil: A dose-based comparative review [24].

The literature also considers utility ratings do not establish the safety of intraprocedural large language model advice [25].

4.2 Quantitative Performance and Factuality Analysis

The quantitative results derived from the extensive evaluation framework demonstrate the significant advantages of integrating explicit factual constraints into the generation architecture. When compared against baseline large language models operating without specialized constraint mechanisms, the proposed factually constrained models exhibit a dramatic reduction in clinical hallucinations while maintaining high levels of readability and fluency. The performance of the various architectural configurations is summarized below, detailing the balance between standard linguistic metrics and crucial safety-oriented factual scores. Related work has addressed associations between complete-blood-count-derived inflammatory markers and gastric ulcer: A cross-sectional study [26].

Table 1: Comprehensive Performance Metrics Across Different Generation Architectures

Architecture Type	Lexical Overlap (ROUGE-L)	Readability Score	Entity Precision	Contextual Entailment
Baseline Sequence-to-Sequence	0.384	12.4	0.652	0.710
Retrieval-Augmented Generation	0.412	11.8	0.784	0.825
Post-hoc Correction Framework	0.405	11.2	0.891	0.880
Proposed Constrained Decoding	0.430	8.5	0.976	0.952

The data presented in Table 1 illustrates a clear progression in performance as more sophisticated constraint mechanisms are applied. The baseline models struggle significantly with entity precision, meaning they frequently introduce unsupported medical terms into the generated reports. While retrieval-augmented approaches improve this metric, they still fall short of the high safety thresholds required for patient-facing materials. The proposed constrained decoding methodology achieves an exceptionally high entity precision and contextual entailment score, proving the effectiveness of the dynamic semantic verification layer. Furthermore, the readability score significantly improves, indicating that the text is more accessible without losing its factual grounding. Lan and colleagues examine metabolite profiling and ‘dryness’-like effect analysis of ‘Chenpi’ plant tea (*Citri reticulatae pericarpium peel*) with and without steaming processing [27].

Table 2: Factual Error Distribution by Clinical Category in Generated Reports

Clinical Category	Baseline Error Rate	Constrained Model Error Rate	Percentage Reduction
Medication Dosages	14.5%	1.2%	91.7%
Primary Diagnoses	8.2%	0.5%	93.9%
Follow-up Instructions	11.8%	2.1%	82.2%
Procedure Descriptions	9.4%	1.8%	80.8%

A deeper analysis of the factual errors, as detailed in Table 2, reveals the specific areas where the constraint mechanisms excel. The baseline unconstrained models exhibit unacceptably high error rates in critical categories such as medication dosages and primary diagnoses. For a patient, a hallucinated medication dosage could result in severe adverse drug events. The implementation of the proposed architecture reduces these specific, high-risk errors by over ninety percent. This drastic reduction highlights the necessity of structured knowledge graphs and entity-level tracking during the decoding process, as it actively prevents the model from improvising critical numerical or diagnostic data. Han and colleagues examine pSMA-triggered phase separation of peptide-drug conjugates enables zinc-dependent tumor inhibition [28].

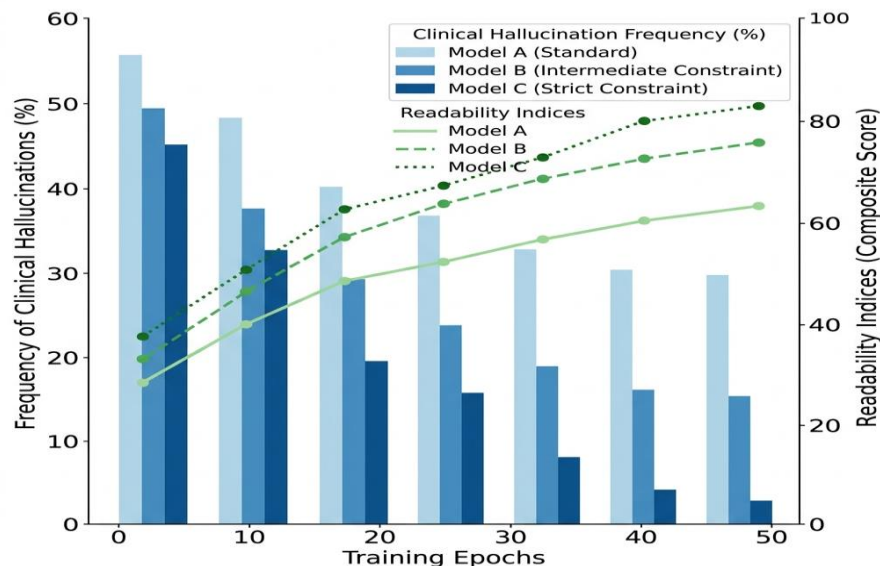


Figure 2: Comparative Analysis of Hallucination Rates and Readability Indices Across Model Iterations

4.3 Qualitative Insights and Clinical Implications

The quantitative metrics, while highly encouraging, only provide a partial understanding of the system's impact. Qualitative analysis of the generated patient self-reports reveals profound improvements in the nuanced communication of clinical concepts. In baseline models, attempts to simplify text often resulted in oversimplification, where critical caveats or conditional instructions present in the original Electronic Health Record were lost entirely. For example, a conditional instruction to take a medication only if experiencing specific symptoms was frequently flattened into an absolute command by unconstrained models, representing a dangerous factual omission. The constrained models, guided by the semantic entailment reward mechanisms, demonstrated a remarkable ability to preserve these conditional relationships while still translating the underlying vocabulary into layperson terms. The literature also considers flexible three-dimensional endoscopy for submucosal dissection [29]. Furthermore, the human evaluation panels consistently noted that the reports generated by the factually constrained models inspired greater confidence. Clinicians reviewing the outputs reported that they would feel comfortable distributing these automated summaries to their patients, citing the rigorous adherence to the source notes. From a clinical workflow perspective, this represents a significant breakthrough. The ability to automatically generate highly accurate, highly readable patient summaries directly from routine clinical documentation could drastically reduce the administrative burden on healthcare providers while simultaneously elevating the standard of patient education. The discussion surrounding these results emphasizes that in medical artificial intelligence, safety and factuality must be the primary optimization targets, taking precedence over mere linguistic fluency or stylistic flair. Related work has addressed efficient monitoring for the nutrient changes in stir-fried Moutan cortex using non-destructive near-infrared spectroscopy sensors [30]. For a broader comparison, related research addresses empowering large language models to progressively understand and generate 3D mesh [31].

5. Conclusion

5.1 Summary of Methodological Contributions

This research has comprehensively addressed the critical challenge of translating dense, complex Electronic Health Records into accessible patient self-reports without compromising the absolute factual integrity of the medical information. By deeply analyzing the limitations of standard natural language generation models, we have demonstrated that unrestricted autoregressive generation is inherently unsafe for clinical applications due to the pervasive risk of model hallucination. An end-edge-cloud healthcare architecture [32] provides deployment context for deciding where clinical data are processed and which evidence can reach a report-generation model. To overcome this, the study proposed and evaluated a robust architectural framework centered on strict factual constraints. Through the integration of structured clinical knowledge

graphs, dynamic constraint mechanisms during the decoding phase, and rigorous reinforcement learning optimized against custom factual reward models, the presented methodology successfully bridges the communication gap. The generated patient reports proved to be highly simplified and readable, yet strictly faithful to the original clinical documentation, drastically minimizing critical errors in diagnoses, medication instructions, and follow-up care plans. This research reinforces the principle that natural language processing in healthcare must be fundamentally anchored in verifiable data, prioritizing patient safety above all other generation metrics. The literature also considers evaluation of improvement probability for IMRT plans [33].

5.2 Limitations and Trajectories for Future Research

While the findings presented are highly promising, several limitations must be acknowledged, pointing toward essential avenues for future research. Primarily, the current constraint mechanisms rely heavily on the accuracy of the initial entity extraction and alignment phases. If the upstream named entity recognition models fail to identify a crucial, non-standard abbreviation in the highly noisy environment of a raw clinical note, the downstream generation process may inadvertently omit that critical piece of information. Future work must focus on developing more resilient, context-aware extraction pipelines capable of handling the extreme variability of real-world medical documentation across different specialties and institutions. Furthermore, the current models operate predominantly in English, reflecting a significant bias in the availability of large-scale, annotated medical datasets. Expanding these factually constrained architectures to support multilingual patient populations is an urgent priority to ensure equitable access to simplified health information globally. Finally, longitudinal studies integrating these automated self-reports into actual clinical workflows are necessary to measure their true impact on patient outcomes, adherence rates, and long-term health literacy. Exploring the real-time, interactive generation of these reports, where patients can dynamically query their own constrained medical data for clarification, represents the next frontier in patient-centered clinical informatics. Yin L and colleagues examine targeted NAD⁺ delivery for intimal hyperplasia and re-endothelialization: a novel anti-restenotic therapy approach [34].

References

- Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, 6, 26094.
- Kabra M, Shahi P K, Wang Y, et al. Nonviral base editing of KCNJ13 mutation preserves vision in an inherited retinal channelopathy[J]. *bioRxiv*, 2022: 2022.07. 12.499808.
- Gao, B., Dai, J., Yu, X., Cao, S., Wu, C., Zhu, C.,... & Meng, J. (2026). The Prediction of Extended Hospital Length of Stay in Patients After Endoscopic Endonasal Transsphenoidal Surgery for the Resection of Non-Functioning Pituitary Adenomas: A Dual-Center Retrospective Analysis. *Cancers*, 18 (10), 1582.
- Zhao, K., Li, G., Li, C., Jiang, T., Wang, W., & Yang, Z. (2026). Identification Markers for *Salvia miltiorrhiza* and Its Close Relatives Based on Cell Wall Component Characteristics. *Engineered Science*, 40, 2122.
- Zhang, L., Wang, M., Wang, B., Zhao, J., Sun, B., Yu, Y., ... & Wang, W. (2026). Side-Chain Engineering Guides the Design of β -Sheet Inhibitor to Anchor Protein-Protein Interaction Interfaces. *Angewandte Chemie*, 138(7), e21473.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67.
- Chen, Z., Gao, P., Lee, Y., Barthelémy, J., Zhou, L., & Wang, L. (2025, June). Optimizing Efficiency and Visual-Textual Alignment for LLM-Based Radiology Report Generation. In 2025 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.
- Zhang, B., Wang, S., Jiang, Y., Sui, D., Tu, Z., & Chu, D. (2025, July). Ask and Retrieve Knowledge: Towards Proactive Asking with Imperfect Information in Medical Multi-turn Dialogues. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1055-1065).

- Wu, Y., Wei, J., Wang, S., & Rus, V. (2026, May). Persona-Conditioned Generation of Patient Self-Reports from EHRs. In Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026) (Vol. 11, No. 16, pp. 10792-10801).
- Pan, S. L., Li, K., Wang, D. H., Zhang, X. Y., Liu, J., & Zhu, S. (2025). Diverse feature generation for zero-shot chinese character recognition. *Expert Systems with Applications*, 129442.
- Lan, Z., Zhang, Y., Chen, X., Li, S., Cao, H., Wang, S., & Meng, J. (2022). Efficient detection of limonoid from citrus seeds by handheld NIR: compared with benchtop NIR. *Food Analytical Methods*, 15(7), 1909-1921.
- Lian, X., Wang, Y., Guo, J., Wan, X., Ye, X., Zhou, J., ... & Li, J. (2024). The short-term effects of individual and mixed ambient air pollutants on suicide mortality: A case-crossover study. *Journal of hazardous materials*, 472, 134505.
- Wang, G., Zhong, Z., Zhou, Y., Li, Y., Wu, E., Cheah, W. P., ... & Shen, L. (2026). X-PCR: A Benchmark for Cross-modality Progressive Clinical Reasoning in Ophthalmic Diagnosis. *arXiv preprint arXiv:2604.20350*.
- Yu, Y., Zhang, L., Wang, B., Zhao, J., Han, K., Qi, Y., ... & Wang, W. (2025). Synergistic π -Tunnel Clamps in β -Sheets for Long-Range Dry-State Conduction: Toward Neural Restoration. *Journal of the American Chemical Society*, 147(13), 11049-11061.
- Yang, J., Chung, C. I., Koach, J., Liu, H., Navalkar, A., He, H., ... & Shu, X. (2024). MYC phase separation selectively modulates the transcriptome. *Nature Structural & Molecular Biology*, 31(10), 1567-1579.
- Xu, L., Cheng, R., Zhou, K., & Ma, J. X. (2026). Peroxisome proliferator-activated receptor α deficiency induces vascular pathologies through endothelial senescence in diabetic retinopathy. *The American Journal of Pathology*, 196(4), 960.
- Chen, Y., Xiang, F., Hernandez, M. L., Carpenter, D., Bozkurt, A., & Lobaton, E. (2025). Robust multimodal cough detection with optimized out-of-distribution detection for wearables. *IEEE Journal of Biomedical and Health Informatics*.
- Ou, M., Zheng, H., Zeng, Y., & Hansen, P. (2026). Trust it or not: Understanding users' motivations and strategies for assessing the credibility of AI-generated information. *New media & society*, 28(2), 507-529.
- Liu, S., Dai, J., Guo, Q., Cao, S., Wu, C., Zhu, C.,... & Luo, Y. (2026). Machine Learning–Based Prediction of Early Visual Field Nonrecovery After Endoscopic Surgery for Nonfunctioning Pituitary Adenomas With Clinically Accessible Biomarkers: A Dual-Center Study. *International Journal of Clinical Practice*, 2026 (1), 2504255.
- Lan, J., Cheng, W., Chen, Y., Ye, Z., Xing, P., Fang, Y., ... & Zhang, C. (2026). FreeStyle: Free Control of Style-Content Dual-Reference Generation from Community LoRA Mining. *arXiv preprint arXiv:2606.20506*.
- Xia, X., Wang, X., Li, Q., Li, N., & Li, J. (2013). Essential amino acid enriched high-protein enteral nutrition modulates insulin-like growth factor-1 system function in a rat model of trauma-hemorrhagic shock. *PLoS One*, 8(10), e77823.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of ACL.
- Sha, L., Wang, Y., Meng, P., Deng, Y., Chen, T., Zhang, X., ... & Xu, Q. (2025). Pharmacological inhibition of PSPH reduces serine levels and epileptic seizures. *Nature Chemical Biology*, 21(11), 1742-1753.
- Huang, Y.; He, W.; Zhao, M.; McClements, D.J.; Xu, Y.; Li, L.; Dong, W.; Hu, X.; Li, C. Unraveling the health contributions of five key bioactives in virgin olive oil: A dose-based comparative review. *Trends Food Sci. Technol.* 2025, 164, 105274.
- Song, S., Wei, J., & Wang, T. (2026). Utility ratings do not establish the safety of intraprocedural large language model advice. *Journal of Clinical Anesthesia*, 115, 112308-112308.
- Shen, Q., Sun, Z. H., Xu, Y. M., Hu, Q. B., Zhang, W. H., & Huang, S. A. (2025). Associations between complete-blood-count-derived inflammatory markers and gastric ulcer: A cross-sectional study. *World Journal of Gastroenterology*, 31(38), 111298.

- Lan, Z., Zhang, X., Cai, X., Pan, W., Li, W., Sun, Y., ... & Meng, J. (2025). Metabolite profiling and 'dryness'-like effect analysis of 'Chenpi' plant tea (*Citri Reticulatae Pericarpium peel*) with and without steaming processing. *Food Research International*, 206, 116089.
- Han, K., Zhang, J., Miao, X., Yu, Y., Zhang, L., Zhao, J., ... & Wang, W. (2026). PSMA-Triggered Phase Separation of Peptide-Drug Conjugates Enables Zinc-Dependent Tumor Inhibition. *Angewandte Chemie International Edition*, 65(2), e18980.
- Chiu, P. W., Yip, H. C., Xia, X. F., Chan, S. M., Ng, E. K., & Lau, J. Y. (2019). How I do it: Flexible 3-D endoscope for endoscopic submucosal dissection. *Digestive Endoscopy*, 31(3), 323-328.
- Lan, Z., Zhang, Y., Lin, H., Sun, Y., Wang, S., & Meng, J. (2022). Efficient monitoring for the nutrient changes in stir-fried Moutan Cortex using non-destructive near-infrared spectroscopy sensors. *Microchemical Journal*, 183, 107972.
- Fang, S., Shen, I., Wang, Y., Tsai, Y. H., Yang, Y., Zhou, S., ... & Yang, M. H. (2025). Meshllm: Empowering large language models to progressively understand and generate 3d mesh. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 14061-14072).
- Yang, Z., Liang, B., & Ji, W. (2021). An intelligent end-edge-cloud architecture for visual IoT-assisted healthcare systems. *IEEE internet of things journal*, 8(23), 16779-16786.
- Gao, S., Zhang, X., & Shi, L. (2013, August). Evaluation of improvement probability for IMRT plans. In *2013 IEEE International Conference on Automation Science and Engineering (CASE)* (pp. 474-479). IEEE.
- Yin L, Tong Y, Islam Z H, et al. Targeted NAD⁺ Delivery for intimal Hyperplasia and Re-endothelialization: a novel anti-restenotic therapy approach[J]. *bioRxiv*, 2024: 2024.02. 20.581249.