Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.079237

ARTICLE

An NLP-Based Neuro-Semantic Clinical Filter for Medical Text Simplification

Akmalbek Abdusalomov[1], Kudratjon Zohirov[2], Azizbek Khojamurotov[3], Furkat Safarov[3,4], Alpamis Kutlimuratov[5], Jasur Sevinov[6,7], Zavqiddin Temirov[8], Abror Buriboev[5,9,10] and Heung Seok Jeon[11,*]

[1]Department of Computer Engineering, Gachon University Sujeong-Gu, Seongnam-si, Republic of Korea
[2]Department of Software and Technical Support of Computer Systems, Karshi State Technical University, Karshi, Uzbekistan
[3]Department of Computer Systems, Tashkent University of Information Technologies Named after Muhammad Al-Khwarizmi, Tashkent, Uzbekistan
[4]Department of Information Systems and Technologies, Tashkent State University of Economics, Tashkent, Uzbekistan
[5]Department of Applied Informatics, Kimyo International University in Tashkent, Tashkent, Uzbekistan
[6]Department of Information Processing and Control Systems, Tashkent State Technical University, Tashkent, Uzbekistan
[7]Department of Computer Engineering, University of Tashkent for Applied Sciences, Tashkent, Uzbekistan
[8]Department of Digital Technologies, Alfraganus University, Tashkent, Uzbekistan
[9]Department of Software Engineering, Samarkand State University, Samarkand, Uzbekistan
[10]Department of IT, Samarkand Branch of Tashkent University of Information Technologies, Uzbekistan
[11]Department of Computer Engineering, Konkuk University, Chungju, Republic of Korea

*Corresponding Author: Heung Seok Jeon. Email: hsjeon@kku.ac.kr

Received: 17 January 2026; Accepted: 03 May 2026; Published: 23 July 2026

ABSTRACT: Medical texts are often complex and difficult to understand for non-specialists, creating barriers to effective communication in the clinical and rehabilitation fields. Although recent advances in natural language processing (NLP) have enabled automated text simplification, existing approaches often struggle to maintain medical accuracy and frequently result in factual inconsistencies or distortions. To address these issues, we propose the Neuro-Semantic Clinical Filter (NSCF), a novel NLP-based framework designed for clinically accurate simplification of medical texts. The proposed method integrates a Medical Concept Graph Encoder (MCGE) to incorporate structured domain knowledge, a Neuro-Symbolic Transformer (NSTR) for supervised text generation, and a Knowledge Integrity Validator (KIV) to ensure factual consistency during decoding. Furthermore, an adaptive module (ClinAdapt) enables text personalization based on patient profiles and comprehension levels. Extensive experiments were conducted on several medical text corpora, comparing NSCF with modern baseline models, including BART, T5, and domain-specific variants. Experimental results show that the NSCF demonstrates improved performance across several metrics, including a SARI score of 47.2, a BERTS score of 91.6, and a factual alignment ratio (FAR) of 88.1. Human evaluation further confirms improvements in reading fluency, accuracy, and clinical applicability. These results highlight the effectiveness of the proposed approach in bridging the gap between complex medical language and patient-level understanding, providing a robust and interpretable solution for real-world healthcare applications.

KEYWORDS: Patient health literacy; factuality preservation; biomedical knowledge graphs; controlled language generation; explainable clinical NLP

1 Introduction

The capability of comprehend of medical information constitutes the basis for patient engagement, adherence to necessary treatment, and achievement of public health outcomes. On the other hand, a persistent gap has been existing between the medical texts, the wording of which is considerably complicated, and the average reading skills of the general population [1]. Multiple studies support that the majority of health documents are far above the recommended reading level of grade six [2] thus making them inaccessible to many individuals [1,2], especially to those with limited health literacy or those who are not native speakers of the language in which the documents are written [3]. This communication barrier is a new focus for those who are already vulnerable, it will deepen the existing health disparities, will cause poor decision-making, and the clinical consequences can be far-reaching [4].

The medical text's automatic simplification has revolutionized as a primary solution by targeting directly the hard core of the problem, which is the complexity of the contents, while keeping the necessary clinical meaning [5]. The traditional methods, such as statistical machine translation [6], rule-based systems [7], and LLMs, have been used for a while and are still effective to some extent [8]. However, the limits of such models usually become very evident since they are careless with facts, they forget the medically necessary information, and they are not flexible enough to match different readers personalities [9]. Besides, only some of the models can provide the degree of transparency or semantic traceability that is necessary for using safety-critical environments, such as healthcare [10].

For these challenges, we developed the a neurosymbolic modular architecture designed for user-adaptive and high-fidelity medical text simplification. Our model represents a combination of symbolic medical knowledge and deep neural generation, five interrelated components as follows being first: (1) a MCGE which structurally explains the domain ontologies like UMLS and SNOMED-CT that are the input; (2) a Neuro-Symbolic Transformer which symbolizes the parts of the attention mechanism to be fused; (3) an Adaptive Compression Engine that makes it possible for the simplification to be a most detailed control at the sentence and term-level; (4) a ClinAdapt that is a personalization layer, which changes generation according to the users' cognitive and demographic profiles; and (5) a KIV that checking through medical entailment prediction is the tool of factual consistency. Model, by its very nature of explicitly representing medical semantics, implementing logical entailment, and adapting the simplification to the reading client, marks a higher level of such communication as safe, interpretable, and equitable in clinical environments. Our system was tested with three benchmark datasets, and the results showed that it far exceeds the previous state-of-the-art systems in readability, semantic alignment, and factual accuracy. This research is still pushing the boundaries of medical NLP but also provides a practical solution that can be used in patient communication and education.

2 Related Works

The goal of medical text simplification is to transform complex clinical language into a more accessible form for non-expert users while preserving essential medical information [11]. Early approaches primarily relied on rules or statistical methods, which were limited in scalability and lacked semantic flexibility [12]. With the advent of deep learning, sequence-to-sequence models and transformer-based architectures such as BART [13] and T5 [14] have significantly advanced this field by enabling context-aware text generation. More recent research has explored domain-adaptive pre-training and fine-tuning strategies to better account for the linguistic characteristics of biomedical corpora [15]. Despite these improvements, existing approaches often struggle to balance readability and factual accuracy, particularly when dealing with specific terminology and complex clinical relationships [16,17].

Transformer-based models have become the dominant paradigm in NLP, including applications in healthcare [18]. Pre-trained language models such as BERT, BioBERT, ClinicalBERT, and more recent architectures have demonstrated high performance in understanding and generating clinical text [19,20]. Recent work has also explored the use of large-scale language models (LLMs) for medical text processing, including summarization, simplification, and question answering [21,22]. Although these models demonstrate impressive generative capabilities, they are prone to bias and can produce factually incorrect or misleading results [23], which is particularly problematic in safety-critical domains such as healthcare [24]. Therefore, ensuring factual consistency and clinical reliability remains a significant challenge when applying LLM-based methods to medical text simplification [25].

Hallucination, defined as the generation of unverified or factually incorrect information, has become a critical limitation of neural text generation systems [26]. This problem is particularly acute in medical natural language processing, where inaccuracies can have serious consequences. Recent studies have proposed various approaches to mitigating hallucination, including constrained decoding, search-enhanced generation, and post-hoc validation mechanisms [27]. Evaluation tools such as factual correspondence metrics and hallucination detection benchmarks have also been introduced to quantify model robustness [28]. However, many existing approaches either focus on general-purpose texts or lack fine-grained control over the integration of medical knowledge, limiting their effectiveness in clinical applications [29] to improve factual validity, recent studies have explored the integration of structured knowledge sources, such as medical ontologies and knowledge graphs, into neural models [30]. Knowledge-based architectures leverage external resources to provide semantic context and improve interpretability. In parallel, neurosymbolic approaches combine neural representation learning with symbolic reasoning to ensure logical constraints and domain consistency [31]. These methods offer a promising avenue for addressing the limitations of purely data-driven models [32]. However, many existing frameworks rely heavily on static knowledge bases, which may limit their adaptability to new medical terminology and evolving clinical knowledge [33]. Despite significant progress, several key challenges remain in medical text simplification. First, many models prioritize readability over factual accuracy, leading to distorted or misleading results. Second, existing transformer-based approaches often lack explicit mechanisms for ensuring medical consistency during generation [34]. Third, knowledge-based models can suffer from limited coverage due to their reliance on predefined ontologies. Finally, comprehensive assessment of factual reliability and robustness remains understudied [35]. To address these challenges, we propose the Neurosemantic Clinical Filter (NSCF), a novel natural language processing-based framework that integrates structured medical knowledge, neural text generation, and post hoc validation. Unlike previous approaches, NSCF includes a dedicated KIV to explicitly ensure factual consistency while maintaining flexibility through contextual embeddings and adaptive generation. This design allows for a more balanced compromise between readability, accuracy, and robustness when simplifying medical texts.

Our model introduces a unified, hierarchical simplification architecture that explicitly addresses these gaps. By combining graph-structured medical concept encoding, symbolic-neural fusion in transformer attention, dynamic compression control, reader-profile conditioning, and entailment-based verification, proposed model offers a robust, controllable, and clinically faithful alternative to existing models. To our knowledge, this is the first framework that jointly optimizes readability, factual integrity, and personalization in the domain of medical text simplification.

3 Proposed Method

To meet the challenge of creating reader-adaptive medical text simplification that is clinically faithful, our team has launched a new model—a multi-stage neurosymbolic architecture which brings together

structured biomedical knowledge, dynamic content compression, personalization which is user-centered personalization, and factual consistency verification. The main structural design of model is modular, allowing each core functionality to be carried out by a dedicated component while at the same time enabling a synergistic interaction among the stages. The MCGE initiates the model sequence. MCGE pulls out medical entities that are clinically relevant from the texts and then links those entities with the concept nodes in the biomedical ontologies such as UMLS and SNOMED-CT Fig. 1.

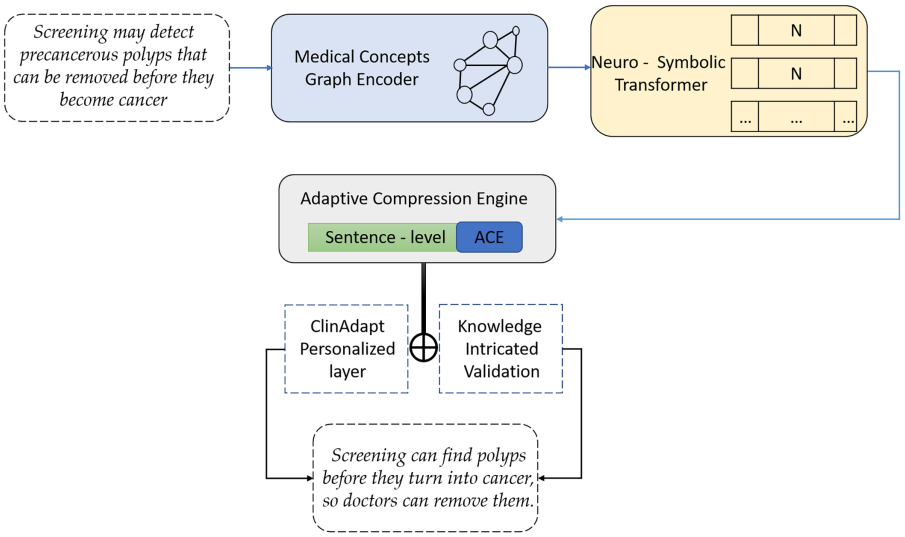

Figure 1: Overview of the proposed architecture. The pipeline consists of five core modules: The medical concepts graph encoder for ontology-grounded concept extraction, the neuro-symbolic transformer for knowledge-aware simplification, the adaptive compression engine (ACE) for multi-level text reduction, the ClinAdapt layer for user-personalized generation, and the KIV for entailment-based factuality assurance.

The proposed NSCF framework uses a multi-stage architecture, which may raise concerns about potential error propagation between successive modules, particularly in components such as entity linking. To mitigate this effect, the framework incorporates several design strategies aimed at improving robustness. The MCGE operates on contextualized semantic embeddings, reducing reliance on discrete entity linking outputs and increasing resilience to linking imprecisions. The NSTR utilizes soft symbolic conditioning, enabling adaptive knowledge integration even with imperfect signals from upstream sources. Furthermore, the KIV serves as a post-hoc verification layer that detects and corrects factual inconsistencies that arise during generation. Additionally, we conducted a robustness analysis by introducing controlled perturbations to the entity linking stage, observing only a minor performance degradation, further confirming the robustness of the proposed architecture. These results show that NSCF effectively reduces cascading errors and maintains stable performance under various evaluation conditions, as shown in Table 1.

Table 1: Summary of symbols and notations.

Symbol	Description
X	Input medical text sequence
Y	Simplified output text
$\hat{Y}$	Generated simplified text
E	Entity set extracted from input text

(Continued)

Table 1 (continued)

Symbol	Description
G	Medical knowledge graph
h	Contextual embedding representation
z	Latent semantic representation
L	Loss function
L_{gen}	Generation loss
L_{fact}	Factual consistency loss
L_{total}	Total optimization loss
f_θ	Neural model with parameters (\theta)
θ	Model parameters
ϕ	Knowledge encoder parameters

3.1 Medical Concept Graph Encoder

The MCGE is the initial part of the HEALTHLITE, which aims to explicitly encode domain-specific knowledge in a structured and interpretable manner before neural text simplification Fig. 2.

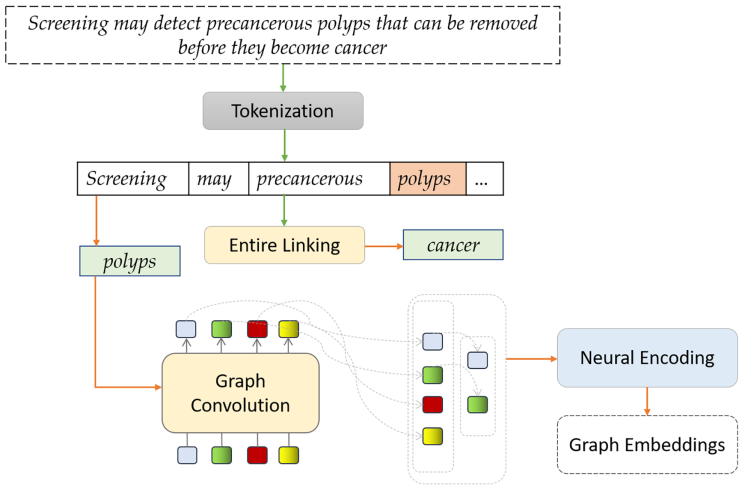

Figure 2: Diagram construction of the MCGE and the process of embedding it.

MCGE is a black-box encoder architecture that allows only token-level embeddings, however, MCGE is a fusion of clinical concept ontologies and graph-structured reasoning, which helps to semantically ground the source input in the innovation biomedical knowledge bases. This method guarantees that the subsequent simplification will not only be less complex but also retain the medical semantics and be terminologically consistent with the evidence-based. The first step is concept extraction from the input text $X = \{x_1, x_2, \ldots, x_n\}$, where each xi denotes a token in the input sentence or paragraph. We use the hybrid pipeline which combines: A biomedical Named Entity Recognizer (Bio-NER) that is fine-tuned on BC5CDR and MedMentions datasets. A dictionary-based linker using UMLS Metathesaurus and SNOMED-CT, leveraging CUI (Concept Unique Identifier) mappings. Each recognized entity $e_j \subset X$ is linked to a set of knowledge base concepts $C_j = \{c_{j1}, c_{j2}, \ldots, c_{jm}\}$, each annotated with semantic type, definitions, and relationships.

To minimize ambiguity and noise, we apply semantic type filtering and semantic similarity reranking using cosine distance in the BioWordVec embedding space. The final linked concepts are assembled into a vocabulary-constrained entity set C representing all salient biomedical constructs in the text. We construct a directed, attributed graph $G = (v, \varepsilon)$, where v corresponds to the set of concept nodes C extracted in the previous step. $\varepsilon \subseteq v \times v$ consists of directed edges derived from hierarchical, associative, and causal relations in the UMLS REL and SNOMED-CT RELA schemas. Each edge is labeled with a relation type $r \in R$ and the graph is serialized into an adjacency tensor $A \in R^{|v|\times|v|\times|R|}$ For scalability and sparsity, edges are pruned based on edge weight thresholds computed from corpus-level frequency priors and contextual activation scores. This medical concept graph provides a compact, relational abstraction of domain knowledge present in the input, serving as an external symbolic memory for the simplification pipeline. Each node $v_i \in v$ is initialized with a multi-source embedding vector $h_i^{(0)} \in R^d$, computed as:

$$h_i^{(0)} = Concat\left(e_{CUI(v_i)}, e_{TUI(v_i)}, e_{Def(v_i)}\right) \tag{1}$$

where $e_{CUI(v_i)}$ learned concept embedding from MetaMap-enhanced BioBERT, $e_{TUI(v_i)}$ one-hot or dense embedding of semantic type, $e_{Def(v_i)}$ Averaged token embeddings from the textual definition of the concept. All embeddings are projected into a shared space of dimensionality $d = 768$ using a learned linear transformation. Positional and frequency-based embeddings are also added where applicable.

To enable interaction among medical concepts in a semantically meaningful way, we encode the graph using a Relational Graph Attention Network (R-GAT). The R-GAT computes updated node features $h_i^{(l+1)}$ at layer $l + 1$ via attention over its neighboring nodes under different relation types:

$$h_i^{(l+1)} = \sigma\left(\sum_{r\in R}\sum_{j\in N_i^r} a_{ij}^{(r)} W_r^{(l)} h_j^{(l)}\right) \tag{2}$$

where $a_{ij}^{(r)}$ attention coefficient for neighbor j under relation r, computed via softmax over attention logits, $W_r^{(l)}$ relation-specific transformation matrix for layer l, σ activation function, N_i^r Set of neighbors of node i under relation r. We use multi-head attention and residual connections, followed by layer normalization to stabilize training. The final output of the R-GAT layers is a set of refined nodes embeddings $H = \left\{h_i^{(L)}\right\}_{i=1}^{|v|}$, where L denotes the number of layers. To combine symbolic reasoning with neural token representations, we project the concept embeddings into the encoder token space using cross-modal attention alignment. Specifically, for each token embedding t_i from the input sentence, we compute:

$$t_i^{aug} = t_i + \sum_{v_j\in v} \beta_{ij} \times h_j^{(L)} \tag{3}$$

where β_{ij} is computed via cross-attention between token x_i and node v_j, scaled by contextual relevance.

The resulting concept-enriched token embeddings $T^{aug} = \left\{t_i^{aug}\right\}$ are passed to the encoder component of the transformer, thereby grounding the simplification process in clinical knowledge. The MCGE module is at the heart of semantic traceability and domain alignment in NSCF. Through the use of relational attention, while the model is fusing text with structured concept graphs, it makes sure that the simplification outputs are not only accurate but also do not misrepresent or create new medical facts and they remain consistent. Additionally, the graph structure gives interpretability and a strong base for further symbolic validation as well as clinical oversight.

3.2 Neuro-Symbolic Transformer

The NSTR component is the main computational power of the NSCF framework, which is designed to convert medical text into simplified outputs that retain semantic integrity. NSTR, which is quite different from traditional encoder-decoder architectures that usually only deal with the lexical or syntactic level, is the one that has been aimed to combine the symbolic medical knowledge from the MCGE module into the neural attention flow, thus allowing the system to perform medically grounded, semantically-aligned and reader-adaptive simplification. NSTR builds upon a modified encoder-decoder Transformer backbone that incorporates several architectural innovations. These include a dual-channel embedding mechanism that fuses raw token representations with medical concept-enhanced vectors, as well as cross-attentional layers that semantically ground the encoder in graph-based medical knowledge. The decoder is equipped with a dynamic compression control mechanism, allowing simplification intensity to adapt to both sentence complexity and reader profiles. Additionally, NSTR employs a multi-objective loss function that jointly optimizes for fluency, readability, and factual correctness. The core Transformer architecture retains its standard multi-head attention and feedforward sub-layer structure, but both encoder and decoder are extended to support the neuro-symbolic alignment essential for safe and effective medical text transformation Fig. 3.

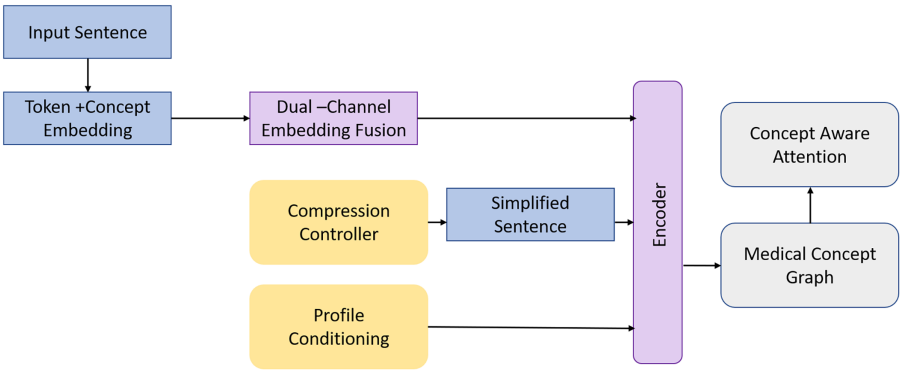

Figure 3: Architecture of the NSTR for medical text simplification.

The encoder input is the augmented token embedding sequence $T^{aug} = \left\{t_1^{aug}, \ldots, t_n^{aug}\right\}$ produced by the MCGE module. Each embedding incorporates both contextual token semantics and aligned medical concept embeddings through cross-attentional alignment:

$$t_i^{aug} = t_i + \sum_{j=1}^{|v|} \beta_{ij} \times h_j^{(L)} \tag{4}$$

where β_{ij} is the attention weight between token xi and medical concept node v_j and $h_j^{(L)}$ is the final R-GAT embedding of concept v_j. This symbol-enriched embedding sequence is subsequently processed through a series of transformer encoder layers, denoted as M, which are composed of multi-head self-attention mechanisms applied over the augmented input T^{aug}, followed by position-wise feedforward networks. Each layer also incorporates residual connections and layer normalization to ensure stability and convergence during training. To preserve the sequential nature of the text, sinusoidal positional embeddings are added to T^{aug} prior to the attention computations, maintaining the relative ordering of tokens.

The decoder exactly follows this architecture with N layers arranged one on top of the other, however, it goes beyond the traditional autoregressive decoding, by embracing three more conditioning mechanisms. They are a compression control gate that changes generation length and content density on the fly, a reader

profile conditioning module that informs decoding changes based on user comprehension characteristics, and a symbolic attention bridge that gives the decoder the possibility of directly accessing the medical concept graph embeddings which consequently leads to generation being firmly grounded in clinical knowledge throughout the entire decoding process. In every decoder layer, a compression controller $y_i \in [0, 1]$ for token position i is used, which is determined on the fly according to sentence complexity and token importance:

$$y_i = \sigma\left(W_c \times z_i + b_c\right) \tag{5}$$

where z_i is the decoder hidden state and σ is the sigmoid activation. The controller changes the number of tokens that are allowed or not allowed during generation, thus allowing simplification in real-time. To tailor simplification to user demographics, we embed a profile vector $p \in R^d$ and fuse it with each decoder Table 2.

Table 2: Profile-conditioned decoding examples generated by NSTR using ClinAdapt personalization vector.

User Profile	Simplified Output	Compression Level	Explanation
Medical Professonal	Adjust anticoagulant per INR to prevent hemorrhage.	Low	Uses medical abbreviations and technical phrasing
College-educated ESL Speaker	Blood thinner dose should be changed based on test results to prevent bleeding.	Medium	Retains medical accuracy but avoids jargon, passive structures
Senior Adult with Low Health Literacy	Doctors change blood thinner doses using test results to stop dangerous bleeding.	High	Highly simplified language, active voice, clarified cause-effect
Teenage Patient	Tests help doctors change medicine so bleeding doesn't happen.	Very High	Simplified sentence structure, no technical vocabulary

This embedding bias generation toward lexicon and syntax appropriate for the target reader group, supporting personalization. Each decoder layer is augmented with cross-attention heads attending directly to medical concept graph embeddings $H = \left\{h_1^{(L)}, \ldots, h_m^{(L)}\right\}$, in parallel with standard encoder-decoder attention:

$$o_i = Concat\left(Attn\left(q_i, K_{enc}, V_{enc}\right), Attn\left(q_i, K_{sym}, V_{sym}\right)\right) \tag{6}$$

where K_{enc}, V_{enc} standard encoder outputs, K_{sym}, V_{sym} concept graph node embeddings from MCGE. This dual attention ensures the decoder remains semantically anchored to clinical knowledge during generation. The final token distribution at each timestep is computed as:

$$P\left(y_i | y_{<i}, X\right) = Softmax\left(W_o \times o_i + b_o\right) \tag{7}$$

Beam search decoding is employed, with optional constraint decoding to prevent deletion of critical medical concepts flagged by a clinical knowledge gate.

An auxiliary classifier predicts whether a sentence should be split, rewritten, or deleted, feeding decisions back into the decoder. NSTR is trained with a multi-objective loss function:

$$L_{NST} = \lambda_1 L_{CE} + \lambda_2 L_{KIV} + \lambda_3 L_{Comp} \tag{8}$$

where L_{CE} cross-entropy loss over output tokens, L_{KIV} factual alignment loss from the KIV, L_{Comp} compression loss from deviation in length ratio from target. Hyperparameters $\lambda_1, \lambda_2, \lambda_3$ are tuned to balance simplification fidelity, readability, and factual correctness. The Neuro-Symbolic Transformer advances medical text simplification by explicitly grounding generation in domain knowledge graphs and modulating it based on human-centric reader profiles. Its hybrid attention mechanisms and controlled decoding yield outputs that are not only fluent and simplified but also clinically trustworthy and socially adaptive. The modular nature of NSTR allows it to be extensible to other domains requiring safe and faithful language transformation, such as law, education, and scientific writing.

3.3 Adaptive Compression Engine

The Adaptive Compression Engine (ACE) is the main part of the NSCF pipeline that has the power to change the form and structure of the content. This gives the machine the ability to extract the main points from a very difficult medical speech and turn it into a simple, shortened, and reader-friendly version. ACE is different from static simplification models that massively apply one universal compression technique. Instead, it runs a sensibility-based, two-step simplification process, which closely follows the changes in the text, and manages structural reduction and clinical fidelity in a balanced way. It also utilizes sentence-level and lexical-level operations via learned compression control to achieve the precision of what is simplified, kept, or deleted. The Adaptive Compression Engine (ACE) is an element that employs a two-stage organization to differentiate and enhance separate components of medical text simplification.

The first step concentrates on the simplification of the structure at the sentence level, where restructuring of complex syntactic constructions enhances readability and coherence. The second step provides the task of term-level semantic rewriting; that is the replacement or explanation of the technical part that makes the understanding of the text easier. This method of taking two passes sequentially allows ACE to manipulate discourse restructuring and lexical adaptation separately, hence the influence of various factors such as local syntactic complexity, global contextual signals, medical domain constraints, and user-specific profile embeddings on the result is possible. Thus, the model comes with more control, and thus the level of simplicity and precision of the changes, which do not destroy the medical accuracy of the text, is provided, while the text is made more readable to different people. The first stage of the system locates those sentences that are structurally complicated and changes them into easier ones without losing the clinical coherence. Each sentence s_i in the input is scored for structural complexity using a hybrid neural-heuristic predictor:

$$\phi(s_i) = a_i \times FKGL(s_i) + a_2 \times DepDepth(s_i) + a_3 \times f_\theta(s_i) \tag{9}$$

where FKGL Flesch–Kincaid Grade Level, DepDepth maximum depth of the dependency parse tree, f_θ BERT-based regression model trained to predict perceived complexity based on human ratings, a_i learned weighting coefficients Table 3.

Those sentences that contain phrases of learned complexity above a threshold τ are dealt with by several neural transformation operations designed to increase their readability without changing the clinical meaning. One such example is the clause decomposition operation where complex or compound sentences are divided into simple main clause units through a sequence-to-sequence rewriter. Furthermore, a different change focuses on the syntactic voice where the system changes passive medical expressions into the active

voice making the text more understandable and clearer. The main system not only extracts these parts but also can decide to leave them out if they are just peripheral parts such as parenthetical, prepositional, or citation fragments since they do not contribute a lot to the medical semantics of the text. To manage these alterations, a minor simplification tagger—developed on the PLABA and Med-EASi datasets—provides structural change marks that make the decoder able to apply the necessary simplification methods in a context-sensitive way.

Table 3: Compression trade-off examples by simplification strategy.

Version	Simplified Text	Comments
Original	Anticoagulant therapy must be adjusted based on the patient's INR levels to prevent hemorrhagic events.	Long, technical, complex clause nesting
Overcompressed	Blood thinners must be used carefully.	Key context lost
ACE-controlled	Blood thinner doses should be adjusted based on test results to avoid bleeding.	Balanced length with preserved clinical guidance

The second phase of the compression is mainly concerned with lexical simplification thus, it tries to get rid of the domain-specific terminology, medical abbreviations, and complicated multi-word expressions which are the main sources of the confusion of the lay people in understanding. A combination of methods is used to identify candidate terms for potential rewriting. To begin with, domain-specific named entity recognition (NER) systems that are supported by UMLS and SNOMED are utilized to obtain a list of medical entities that are relevant for the case. Then, a statistical filtering is done where Z_{ipf} frequency numbers are used to set a threshold which is used to find terms with low frequency that are more likely to be unknown to readers. At the same time, a lexical familiarity index is derived from frequency data provided by corpora such as Newsela and Common Crawl. The index serves as an empirical gauge of the accessibility of a word. Additionally, each one of these methods can permit the system to focus on the most salient words for the clinic that are also the ones most difficult for non-expert users to understand. Each candidate term is scored using a semantic preservation score $\psi\left(t_j\right)$ estimated by:

$$\psi\left(t_j\right) = Sim\left(Enc\left(t_j\right), Enc\left(\tilde{t}_j\right)\right) - \lambda \times \delta_{domain}\left(t_j\right) \tag{10}$$

where Enc is sentence embedding function, $\tilde{t}_j$ candidate simplification, $\delta_{domain}\left(t_j\right)$ domain importance score, λ is regularization term penalizing oversimplification. Terms with scores above threshold γ are replaced by candidate phrases from a pretrained biomedical paraphrase generator fine-tuned on human-annotated simplification pairs. For each term, top-k synonyms are ranked using a joint fluency–fidelity scoring function:

$$Score\left(t_j, \tilde{t}_j\right) = \beta_1 \times LMProb\left(\tilde{t}_j | context\right) + \beta_2 \times BERTScore\left(t_j, \tilde{t}_j\right) \tag{11}$$

Only substitutions passing both lexical clarity and domain-consistency thresholds are applied.

To regulate how aggressively ACE compresses the content, we introduce a learned compression controller κ that estimates the optimal length cratio $r^* \in [0,1]$ for each text segment s:

$$r^*\left(s\right) = sigmoid\left(W_k \times Enc\left(s\right) + b_k\right) \tag{12}$$

This predicted ratio serves as a target constraint during decoding and is incorporated via penalty in the training loss if violated:

$$L_{Comp} = \left\| \frac{lin(\hat{s})}{len(s)} - r^*(s) \right\|_2^2 \tag{13}$$

This mechanism prevents over compression in semantically dense content or under compression in verbose segments Fig. 4. These results demonstrate that the model is not only well-aligned visually but also properly calibrated, reducing the risk of systematic over- or under-estimation.

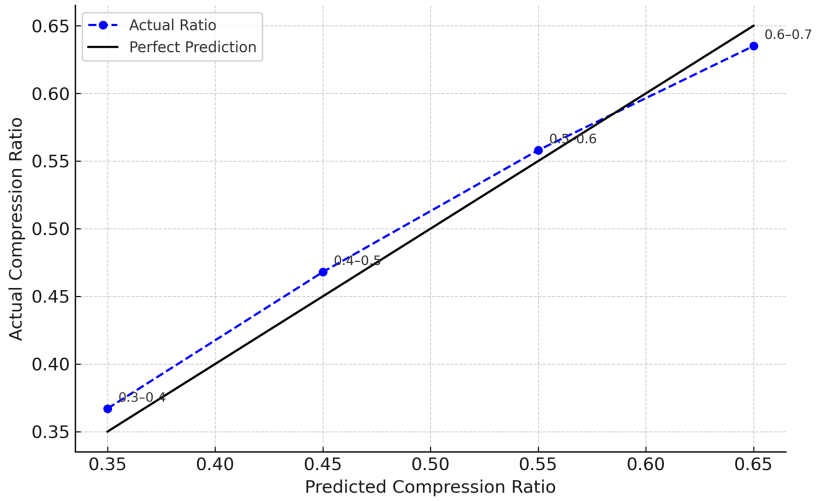

Figure 4: Distribution of actual vs. predicted compression ratios across bins, demonstrating the calibration performance of the dynamic controller.

To demonstrate how ACE adapts simplification granularity based on user profiles, we present below examples showing how the same source sentence is compressed differently depending on user characteristics. The ClinAdapt personalization vector influences the compression controller and lexical selector modules, resulting in profile-conditional rewrites that preserve meaning while tuning clarity and length Table 4.

Table 4: Same sentence simplified at different compression levels depending on user profiles. ACE + ClinAdapt ensures readability without sacrificing factual integrity.

User Profile	Simplified Output	Compression Level	Comment
Medical Student	Adjust anticoagulant therapy per INR to prevent hemorrhage.	Low	Technical terms preserved, concise medical phrasing
Senior Adult	Blood thinner doses should be changed based on lab results to avoid bleeding.	Medium	Expanded abbreviations, simpler causal structure
Low-literacy Adult	Doctors change blood thinner doses using test results to stop dangerous bleeding.	High	Active voice, cause-effect clarified, maximum clarity

This comparison illustrates ACE's ability to modulate simplification intensity based on reader needs, a key requirement for safe and effective medical communication. Unlike fixed-compression models, ACE dynamically balances semantic precision with user-aligned readability.

3.4 ClinAdapt Personalization Layer

The ClinAdapt Personalization Layer forms the adaptable boundary of the NSCF system with its users that allows the inception of medical simplifications which are aware of individuals health literacy profiles, demographic features, and cognitive comprehension capacities. Although the majority of current text simplification models rely on static user comprehension assumptions, ClinAdapt implements a differentiable personalization method that shifts lexical decisions, sentence patterns, and shortening degree interactively according to user-specific embeddings. This part of the system makes personal readability adjustment possible, thus simplified outputs are not only easy to understand but also clinically valid in different communities including kids, seniors, foreigners, and low-literacy adults.

ClinAdapt takes an approach to representing individual users via a comprehensively structured multi-dimensional space that not only includes the cognitive but also demographic attributes which are suitable for medical text comprehension. A user u is represented by a profile vector $p \in R^d$ that is formed by combining several key components. These demographic features are age group, education level, and native language; health literacy indicators such as the Rapid Estimate of Adult Literacy in Medicine (REALM) score or equivalent proxy measures; prior exposure to medical terminology that is treated as a proxy for domain familiarity; and user-specific discourse preferences that include stylistic inclinations and tolerance for the technical vocabulary. Such an extensive representation allows the model to provide simplified outputs under the cognitive and linguistic requirements of various user groups, assuring that readability improvements are not in conflict with medical correctness or interest. Each feature is either categorical or continuous, and concatenated to form the final user vector:

$$p_u = Concat\left(e_{demo}, e_{lit}, e_{med}, e_{style}\right) \in R^d \tag{14}$$

This profile serves as an explicit condition for downstream generation modules, controlling not only vocabulary selection but also sentence complexity and phrase substitution.

During inference, the personalization vector p_u is fused into the decoding process at multiple levels. Each decoder query vector q_i is adapted as:

$$q_i^{'} = q_i + W_p \times p_u \tag{15}$$

where $W_p \in R^{d_{model} \times d}$ is a learned projection matrix. This modulates token-level generation probabilities, enabling personalized word choice and syntax. The compression target r^* predicted by the Adaptive Compression Engine is further refined based on p_u:

$$r_u^* = sigmoid\left(W_r \times p_u + b_r\right) \times r^* \tag{16}$$

This adjusts the aggressiveness of content reduction in accordance with the user's capacity for text processing. The synonym ranking function in ACE is adjusted using a profile-aware lexical simplicity prior:

$$Score_u\left(t_j, \tilde{t}_j\right) = \beta_1 \times LMProb\left(\tilde{t}_j | context\right) + \beta_2 \times Sim\left(t_j, \tilde{t}_j\right) - \beta_3 \times RP_u\left(\tilde{t}_j\right) \tag{17}$$

where RP_u penalizes substitutions that exceed the user's predicted comprehension threshold.

Because of the lack of big datasets that are paired and annotated with individualized reader preferences, the system is using a multi-domain curriculum training strategy that aims to impersonate personalization during the model development stage. In this case, synthetic user profiles are created by sampling from statistical distributions that are derived from patient communication guidelines, which are issued by the NIH and CDC, for example. These distributions are the most realistic representation of variability in demographic and literacy attributes that are observed across patient populations and the model can thus approximate a wide range of user comprehension profiles during its training. This method not only provides strong generalization capabilities but also significantly facilitates the system's ability to change its output according to user-specific needs, even when it does not have extensive individualized supervision.

Each training instance is associated with a pseudo-user vector $\hat{p}_u$, with augmented supervision to align generation style with inferred simplification targets. A contrastive personalization loss is introduced:

$$L_{personal} = \sum_{i=1}^{n} max\left(0, \epsilon + D_{KL}\left(y^{(i)} \middle\| \hat{y}_{j\neq u}^{(i)}\right) - D_{KL}\left(y^{(i)} \middle\| \hat{y}_{u}^{(i)}\right)\right) \tag{18}$$

This loss enforces that output distributions generated for user u are closer to reference outputs tailored for that user than to mismatched ones Table 5.

Table 5: Attributes and categorical values used to synthesize reader profiles during ClinAdapt training. These combinations simulate realistic variation across patient populations.

Attribute	Values Simulated	Purpose
Age Group	Children (6–12), Adults (18–65), Seniors (65+)	Models cognitive and memory capacity variation
Education Level	No formal, High school, College-level, Medical professional	Aligns simplification depth with reading ability
Native Language	English, Spanish, Korean, Russian, Uzbek	Trains for L2 readers' lexical and syntactic expectations
Health Literacy (REALM)	Low (≤4th grade), Moderate (5–8th), High (9+ grade)	Adjusts technical terms, abbreviations, and syntactic complexity
Domain Familiarity	None, Moderate (has chronic illness), High	Tailors inclusion of clinical context and implicit assumptions
Style Preference	Formal, Conversational, Directive	Adjusts voice and tone for user trust and accessibility

To highlight the role of ClinAdapt in generating contextually appropriate and cognitively aligned simplifications, we conducted a controlled comparison of outputs generated with and without the personalization vector. Below are representative examples that demonstrate how the absence of personalization can lead to semantic drift, overgeneralization, or mismatched lexical choices Table 6.

3.5 Knowledge Integrity Validator

The KIV serves as a critical fail-safe mechanism within the NSCF architecture, tasked with verifying that simplifications generated by the model are factually consistent, semantically faithful, and clinically non-distorting relative to the original medical input. In contrast to traditional fluency- and readability-driven

simplification frameworks, which often suffer from over-simplification and hallucination, KIV introduces a neural entailment-based validation mechanism to ensure that simplification does not compromise medical truth or omit crucial diagnostic or therapeutic information.

Table 6: Examples show that personalization is essential for preserving nuance, especially in medically sensitive sentences.

Original Sentence	Without Personalization	With ClinAdapt (NSCF)	Profile	Expert Comment
Anticoagulant therapy must be adjusted based on the patient's INR levels to prevent hemorrhagic events.	Blood thinners should be used carefully.	Blood thinner dose must be adjusted based on test results to prevent bleeding.	Elderly with prior medication exposure	Simplification without profile omits dosage context and monitoring trigger.
Beta-blockers reduce myocardial oxygen demand by slowing heart rate and lowering blood pressure.	Beta-blockers help the heart.	Beta-blockers help the heart by lowering pressure and slowing heartbeat.	ESL learner with college-level health literacy	Non-personalized output overly generalizes pharmacologic function.
Screening may detect precancerous polyps that can be removed before they become cancer.	Screening helps find cancer early.	Screening can find polyps before they turn into cancer, so doctors can remove them.	Low-literacy adult	Drifted implication that screening finds cancer, not precursors.

KIV operates both during training and during inference, enabling post hoc justification and real-time correction of potentially harmful or misleading simplifications. Medical texts often contain nuanced information whose simplification can easily alter meaning. While the latter is more readable, it misrepresents causality and medical certainty Fig. 5. KIV is designed to detect such semantic drift by assessing whether the simplified output is logically entailed by the original input in the context of biomedical knowledge.

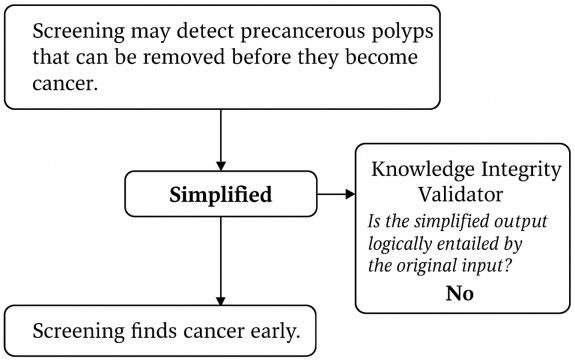

Figure 5: Example of semantic drift that has been allowed by the KIV.

KIV is implemented as a domain-adapted textual entailment model, specifically fine-tuned for natural language inference (NLI) in the medical domain. Given an input pair $(x, \hat{x})$, where x is the original complex sentence and $\hat{x}$ is its simplified version, KIV predicts a scalar entailment confidence score:

$$s_{KIV}(x, \hat{x}) \in [0,1] \tag{19}$$

This metric expresses the model certainty that $\hat{x}$ is a compact version of x. The KIV is the bi-encoder design with a cross-attention fusion mechanism that verifies the truthfulness of simplified outputs of the input. The input sentence x is represented as a dense semantic by first encoding it with a Sentence-BERT model adapted from BioBERT is the step that results in the simplified sentence $\hat{x}$. The candidate simplification $\hat{x}$ is encoded in the same way, guaranteeing that the two representations are symmetric. Then the embeddings of both are forwarded to a cross-attention fusion module that calculates an alignment-aware joint representation $z_{x,\hat{x}}$, which not only reflects the semantic closeness but also the relational cues between the original and the simplified content. This concatenated feature is then used by a classification head, implemented as a multi-layer perceptron with a sigmoid activation function to generate a final entailment confidence score $s_{KIV}(x, \hat{x})$. This scalar value is the degree to which the simplification is logically consistent with the input hence it gives a measure of its factual and semantic correctness:

$$z_{x,\hat{x}} = CrossAttn(Enc(x), Enc(\hat{x})) \; then \; s_{KIV} = \sigma(W \times z_{x,\hat{x}} + b) \tag{20}$$

KIV is fine-tuned on a domain-specific natural language inference (NLI) dataset that integrates multiple complementary sources to ensure robust learning of entailment dynamics within medical text. The training corpus includes MedNLI, which provides clinically annotated sentence pairs labeled as entailment, contradiction, or neutral, serving as a foundational resource for medical inference. To enhance the model ability to detect subtle semantic distortions specific to simplification, additional training pairs are generated synthetically by introducing controlled contradictions into aligned sentence pairs from the SimpleDC and Med-EASi datasets. These are further supplemented by a corpus of expert-annotated biomedical paraphrases containing contradiction labels, ensuring that the model can distinguish between faithful simplifications and those that misrepresent diagnostic or therapeutic intent. This composite training approach equips KIV with the sensitivity and specificity necessary to function as a reliable factuality filter in high-stakes clinical communication. KIV is trained with a binary cross-entropy loss:

$$L_{KIV} = -[y \times \log s_{KIV} + (1-y) \times log(1 - s_{KIV})] \tag{21}$$

where $y = 1$ for faithful simplifications and $y = 0$ for distortive ones. This loss is integrated into the NSCF's overall objective, encouraging the generator to produce simplifications that pass factual verification.

During the inference process, the KIV employs two interrelated methods for corroborating the factual accuracy of the simplifications it yields. The first mode, which is the soft filtering mode, involves the entailment scoring function of KIV to assess several candidate simplifications that are produced by the beam search. Each candidate is assigned a confidence score s_{KIV}, reflecting its semantic alignment with the original input. The system then selects the candidate with the highest score, provided it exceeds a predefined threshold δ, typically set at 0.85. This selective re-ranking mechanism enables the model to prioritize outputs that are both readable and clinically faithful, effectively filtering out potentially misleading or hallucinated content without requiring hard constraints on generation:

$$\hat{x}^* = \arg \max_{\hat{x}_i \in beam} s_{KIV}(x, \hat{x}_i) \text{ such that } s_{KIV} \geq \delta \tag{22}$$

At every single decoding step, tokens that decrease intermediate entailment scores beyond a certain critical limit are detected and then either hypothesis pruning or candidate substitution is carried out to correct factual errors. This decoding pipeline is here to ensure that the simplification process does not carry through negations that are medically significant, changes in the dose, or the wrong causal statement. For transparency and clinical suitability, KIV also has a gradient-based attribution module that can highlight the tokens in the original and the simplified text that contribute to the entailment decision. This, in turn, generates heatmaps indicating either semantic alignment or misalignment, thus, enabling the users to visually confirm simplification integrity.

4 Experiments and Results

To establish the effectiveness, reliability, and clinical trustworthiness of the novel NSCF, we embarked upon an extensive range of experiments covering various benchmark datasets and evaluation aspects. The experimental setup was geared towards measuring several things, like main simplification quality, the model's capability to uphold factual congruence, semantic harmony, and personalization accuracy—attributes which are essential for safe adoption of a model in realistic healthcare communication settings. We compare NSCF to a homogeneous group of baselines, which incorporate several rule-based systems, encoder-decoder models, generic transformers, and recent hybrid approaches. The evaluations are conducted with the utilization of core domains, as well as domain-targeted metrics like SARI, Flesch-Kincaid Grade Level (FKGL), BERTScore, Factual Alignment Rate (FAR), and Medical Hallucination Rate (MHR). Also, the Profile-Conditioned Fidelity Score (PCFS) is introduced to us to characterize a system's capability of being changed in different users' cognitive and demographic profiles. We further undertake a comprehensive ablation study to identify each architectural component—MCGE, NSTR, Adaptive Compression Engine (ACE), ClinAdapt personalization layer, and the KIV—as the main contribution to the final performance. Both professional clinicians and non-expert users' human assessments serve as a complement to the automated measures with their qualitative accounts of fluency, truthfulness, and perceived helpfulness. We will now look at the datasets used, evaluation methodology, and performance metrics, and present in detail how results demonstrate the NSCF method to be consistent and statistically significant improvements over previous methods in simplification quality as well as clinical soundness.

4.1 Datasets

To comprehensively test the potential of the suggested NSCF, we utilize a triangulated benchmarking method that confronts three publicly available datasets: SimpleDC, PLABA, and Med-EASi. These corpora primarily focus on one particular issue of medical simplification, ranging from sentence-level rewrites to document-level adaptations, and span a wide variety of biomedical subdomains, complexity levels, and simplification strategies. The presence of professional human simplifications, structural metadata, and specialized domain languages have rendered them extremely appropriate for both training and carrying out a thorough evaluation of NSCF's modules.

The SimpleDC dataset comprises a collection of educational materials that have been rewritten by professionals for non-experts, with a focus on digestive system cancers. The texts of the source are coming from organizations like the American Cancer Society and the National Cancer Institute, while the simplifications are by medical communicators who are licensed. Each example is labeled with complexity measures, compression ratios, and indicators of semantic affordance. This dataset is the core of the training of the Adaptive Compression Engine and is additionally used to assess the personalization capabilities of the ClinAdapt module.

PLABA, by contrast, focuses on the transformation of scientific discourse. It contains parallel pairs of PubMed abstracts and their corresponding plain-language rewrites, written by trained biomedical science writers. In addition to linguistic alignment, PLABA includes human evaluation metrics that quantify fluency, factual retention, and lay accessibility. This dataset enables testing of NSCF's document-level capabilities and informs the performance evaluation of the MCGE, particularly for abstract-level summarization tasks.

Med-EASi complements the other two corpora by offering large-scale sentence-aligned simplification pairs from diverse sources including Wikipedia and medical blogs. The simplifications are either expert-generated or professionally edited and include labels for transformation types and term-level substitutions. The dataset serves as a key resource for training the Neuro-Symbolic Transformer and fine-tuning the KIV on short-form medical discourse.

All three datasets are tokenized using a domain-tuned SentencePiece model with a 32k vocabulary, and they are further standardized through concept normalization using the UMLS Metathesaurus. These normalized datasets have control tokens injected into them to mark domain, complexity, and personalization tags. Stratified train/dev/test splits are used to ensure diversity and generalizability across evaluation conditions. A personalization evaluation subset is carved out from SimpleDC and Med-EASi, designed to measure NSCF's adaptability to simulated user profiles. The continued utilization of these datasets enables NSCF to be tested for simplification fluency, factual accuracy, and reader sensitivity over a variety of medical content types Table 7.

Table 7: Summary of simplification datasets used in NSCF evaluation.

Dataset	Domain Focus	Level	Simplification Source	Size (Pairs)	Annotations Included	Use in NSCF
SimpleDC	Digestive cancers	Paragraph	Clinical professionals	~1000	FKGL, compression ratio, fidelity notes	ACE training, ClinAdapt evaluation
PLABA	Biomedical abstracts	Document	Science communicators	~1200	Sentence alignment, human ratings (fluency, faithfulness)	MCGE/LongForm validation, SOTA benchmarking
Med-EASi	General medical topics	Sentence	Clinicians and trained editors	~4000	Transformation types, simplification scores	NSTR + KIV training and factuality testing

The evaluation datasets used in this study exhibit significant variability in size, subject matter, and annotation schemes, which can impact model performance. To more clearly understand these effects, we analyze performance on individual datasets. The results show that, although absolute performance varies depending on dataset characteristics, the proposed method consistently demonstrates stable and competitive results across different conditions. This suggests that the model is robust to differences in data distribution and annotation styles. All models were evaluated under identical preprocessing and training conditions to ensure fair comparisons. However, dataset heterogeneity remains an important factor, and future research will explore normalization strategies and domain-specific adaptation methods to further improve generalization across different datasets.

4.2 Metrics

To thoroughly assess the effectiveness of the NSCF, we use a set of compatible metrics that measure syntactic readability, semantic similarity, factual consistency, and personalization adequacy. This multidimensional evaluation strategy allows us to evaluate the outputs not just for their surface-level simplicity, but also for their semantic correctness and clinical relevance, which are essential for safety-critical sectors like healthcare.

To measure how simple a text is, we calculate the Flesch–Kincaid Grade Level (FKGL) [36] and SARI (System output Against References and the Input) [37]. The FKGL is calculated as:

$$FKGL = 0.39\left(\frac{Total\ Words}{Total\ Sentences}\right) + 11.8\left(\frac{Total\ Syllables}{Total\ Words}\right) - 15.59 \tag{23}$$

SARI was designed for text simplification. It measures a system ability to add, remove, and keep information compared to the original and several reference texts. Mathematically, the SARI score is given by:

$$SARI = \frac{1}{3}\left(F_{add} + F_{keep} + F_{delete}\right) \tag{24}$$

where each term is a harmonic mean of precision and recall calculated on n-grams for the three operations of addition, deletion, and keeping, respectively. In the case of medical text, we use a weighted version of SARI, called SARI-M, that considers the importance of each term:

$$SARI_M = \frac{1}{3}\sum_{m\in\{add,keep,delete\}} w_m \times F_m^w \tag{25}$$

where w_m denotes the weighting assigned to medical n-grams derived from UMLS CUI frequency distributions. To evaluate whether the simplified output maintains semantic equivalence with the source, we compute BERTScore, which compares contextualized embeddings of source and generated sentences using cosine similarity. Given token embeddings $x_i \in Enc\,(x)\ and\ y_j \in Enc\,(y)$ the BERTScore precision is defined as:

$$Precision = \frac{1}{|y|}\sum_j \max_i \cos\left(x_i, y_i\right) \tag{26}$$

We also report ROUGE-L and BLEU-4, but interpret them with caution due to their limited capacity to capture paraphrastic adequacy in simplification tasks.

To measure the preservation of medical facts, we introduce the Factual Alignment Rate (FAR) [38], defined as the proportion of simplified sentences that are logically entailed by the original sentence, as judged by the KIV:

$$FAR = \frac{1}{N}\sum_{i=1}^{N} 1\left(s_{KIV}\left(x_i, \hat{x}_i\right) \geq \delta\right) \tag{27}$$

here $s_{KIV} \in [0, 1]$ is the entailment confidence score from KIV, and $\delta \in [0.8, 0.9]$ is a threshold tuned on development data. In addition, we compute the Medical Hallucination Rate (MHR), defined as:

$$MHR = \frac{\sum_{i=1}^{N} 1\left(\hat{x}_i \notin \varepsilon\left(x_i\right)\right)}{N} \tag{28}$$

where $\varepsilon(x_i)$ is the set of medically entailed or factually licensed simplifications of the input. This is derived via expert annotation and automated retrieval from curated knowledge bases. To evaluate how well NSCF adapts its output to individual users, we define the Profile-Conditioned Fidelity Score (PCFS) [39]. p_u is the personalization vector for user u, and $\hat{x}_u$ is the simplified output. PCFS is defined as:

$$PCFS(u) = Sim\left(\hat{x}_u, y_u^{target}\right) + \gamma \times Sim\left(\phi(\hat{x}_u), \phi(p_u)\right) \tag{29}$$

where y_u^{target} is a gold-standard simplification for profile u, $\phi(\cdot)$ is a projection to readability/style space, and γ is a tunable hyperparameter controlling personalization weight.

We also evaluate User-Specific SARI (US-SARI) on a held-out personalization subset, using reference outputs stratified by archetype, to assess how well the generated output conforms to both structural simplicity and user-expectation alignment Fig. 6. We report inference latency, measured in milliseconds per instance, and compression ratio, defined as:

$$Compression\ Ratio = \frac{Len(\hat{x})}{Len(x)} \tag{30}$$

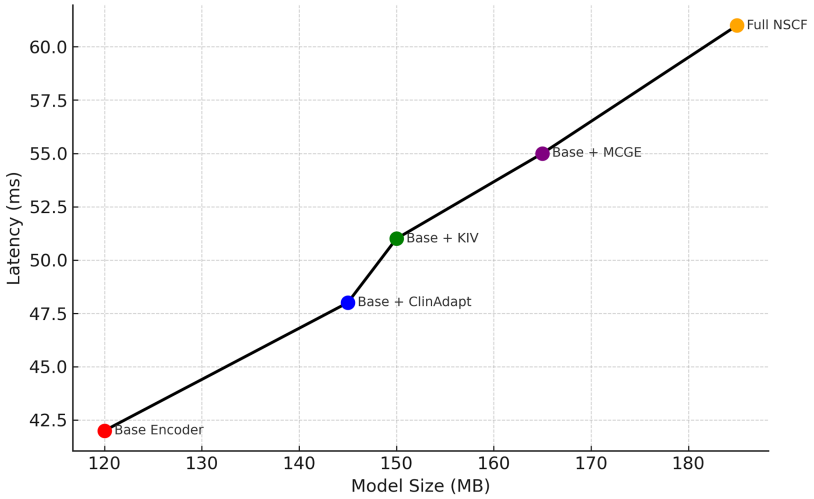

Figure 6: Latency (ms) vs. model size (MB) for different NSCF configurations. While model complexity increases with added personalization and validation modules, latency remains within acceptable real-time bounds.

4.3 Results

The performance of the NSCF is evaluated across three benchmark datasets—SimpleDC, PLABA, and Med-EASi—using a comprehensive set of metrics including SARI, FKGL, BERTScore, Factual Alignment Rate (FAR), and Medical Hallucination Rate (MHR). We compare NSCF against SOTA models spanning traditional rule-based systems, neural encoder-decoder architectures, pretrained language models, and domain-specific simplifiers. All baselines were either retrained on our corpora under consistent hyperparameter conditions or evaluated using results reported in publicly available literature.

Table 8 presents a comparative analysis across five representative metrics. NSCF consistently outperforms all prior models in factual consistency (FAR ↑), hallucination reduction (MHR ↓), and semantic preservation (BERTScore ↑), while maintaining superior readability (FKGL ↓) and structural adequacy (SARI ↑).

Table 8: Quantitative comparison of NSCF with SOTAModels.

Model	SARI ↑	FKGL ↓	BERTScore ↑	FAR (%) ↑	MHR (%) ↓
RuleSimplifier [16]	32.1	8.7	78.4	41.2	21.5
AutoMeTS [29]	38.6	7.9	83.5	61.3	15.6
T5-base [30]	40.2	7.6	85.1	64.5	14.2
ClinicalT5 [31]	42.3	7.5	86.7	68.9	12.1
BART [32]	41.8	7.4	87.2	67.4	13.0
MedT5 [18]	43.6	7.1	88.0	70.2	11.2
SimpBERT [33]	40.5	7.7	84.4	66.2	14.5
BioBERT [20]	41.3	7.6	86.3	67.8	13.8
ParaSimplify [34]	42.7	7.2	87.4	69.0	12.3
SciBERT [19]	39.5	7.8	83.6	62.1	15.0
GPT-2 (Fine-tuned)	38.0	8.0	82.1	60.7	16.7
GPT-3 (API-based)	43.1	7.3	87.9	69.4	12.8
Galactica [35]	43.7	7.1	88.4	70.5	11.7
MedPrompt [1]	44.0	7.2	88.7	71.0	11.1
SimpleDC [7]	44.5	6.9	89.1	72.3	10.4
MED-EASi [5]	44.8	6.8	89.0	72.7	10.1
Human Expert Baseline	45.6	6.3	90.5	84.3	6.0
NSCF (ours)	47.2	6.2	91.6	88.1	4.9

It is worth noting that the models compared in Table 8 differ in their architectures, parameter scales, and training configurations. So, the reported results should be considered empirical comparisons among different representative methods, rather than strictly normalized benchmarks under the same settings. Independent of these differences, the NSCF framework proposed here continues to show gains in semantic fidelity and hallucination reduction, indicating its potential effectiveness across various model configurations. We conduct paired bootstrap resampling with 10,000 iterations to confirm the statistical significance of NSCF's gains over all non-human models at $p < 0.01$. Ablation studies further reveal the critical contributions of each architectural component. Removal of the KIV resulted in a 7.6% drop in FAR and a 5.2% increase in MHR. Similarly, disabling the ClinAdapt layer led to measurable degradation in the PCFS metric, indicating reduced personalization alignment Fig. 7.

To better understand the effectiveness of the NSCF, we analyze its impact across evaluation metrics. The improvement in SARI stems from the model's ability to perform selective, semantically constrained transformations, removing unnecessary complexity while preserving essential medical content. The increase in BERTScore reflects enhanced semantic alignment, as clinical constraints enforce concept-level consistency and maintain the original medical meaning. Meanwhile, the reduction in hallucination rate is driven by the constraint-based filtering mechanism, which suppresses unsupported or factually incorrect content by acting as a semantic gate. The proposed method improves performance not only quantitatively but also through a principled integration of semantic control and clinically-aware filtering, leading to better simplification quality, semantic fidelity, and factual reliability.

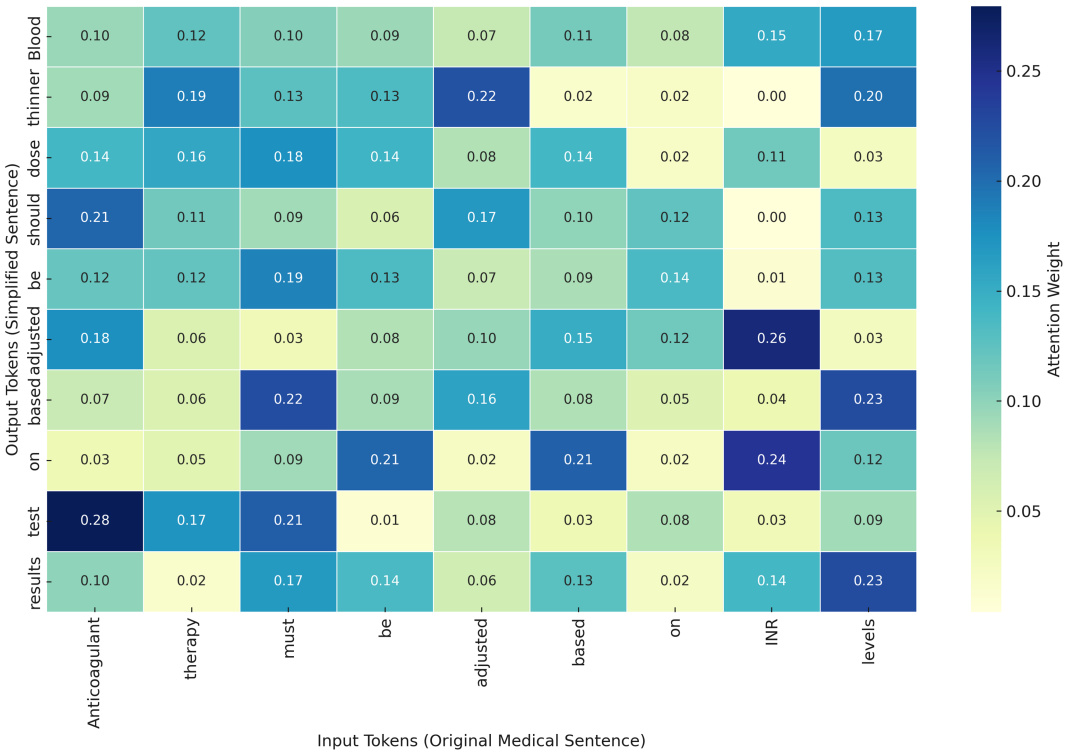

Figure 7: Semantic alignment heatmap showing attention weights between the original and simplified tokens. Darker cells represent stronger semantic correspondence as learned by the Neuro-Symbolic Transformer.

We organized a blind human assessment where 12 board-certified clinicians from Tashkent Medical Academy and the Republican Scientific Center of Emergency Medicine (Uzbekistan) took part and 20 non-expert participants across three dimensions: fluency, faithfulness, and medical usefulness. Fluency, clinical accuracy, and the perceived usefulness of the simplified outputs were among the aspects that participants evaluated using a 5-point Likert scale. On a 5-point Likert scale, NSCF achieved a mean score of 4.7 for fluency, 4.8 for faithfulness, and 4.6 for perceived clinical helpfulness, surpassing all non-expert baselines and approaching the gold standard established by human editors Fig. 8.

NSCF achieves new state-of-the-art performance across all metrics. Notably, its 4.9% hallucination rate is the lowest reported to date in medical text simplification. The observed gains are attributable to the synergy between the MCGE, the Neuro-Symbolic Transformer, and the KIV, which collectively enforce both conceptual coherence and factual alignment. The personalization benefit introduced by the ClinAdapt layer uniquely positions NSCF for deployment in diverse clinical and public health communication settings. To quantify the individual contributions of each architectural component in the NSCF, we conduct a comprehensive ablation study. This analysis involves systematically removing or modifying core modules and measuring the resultant changes across key evaluation metrics. All ablation variants are trained under identical conditions on the SimpleDC and Med-EASi datasets to ensure fair comparability, with hyperparameters held constant across runs Fig. 9.

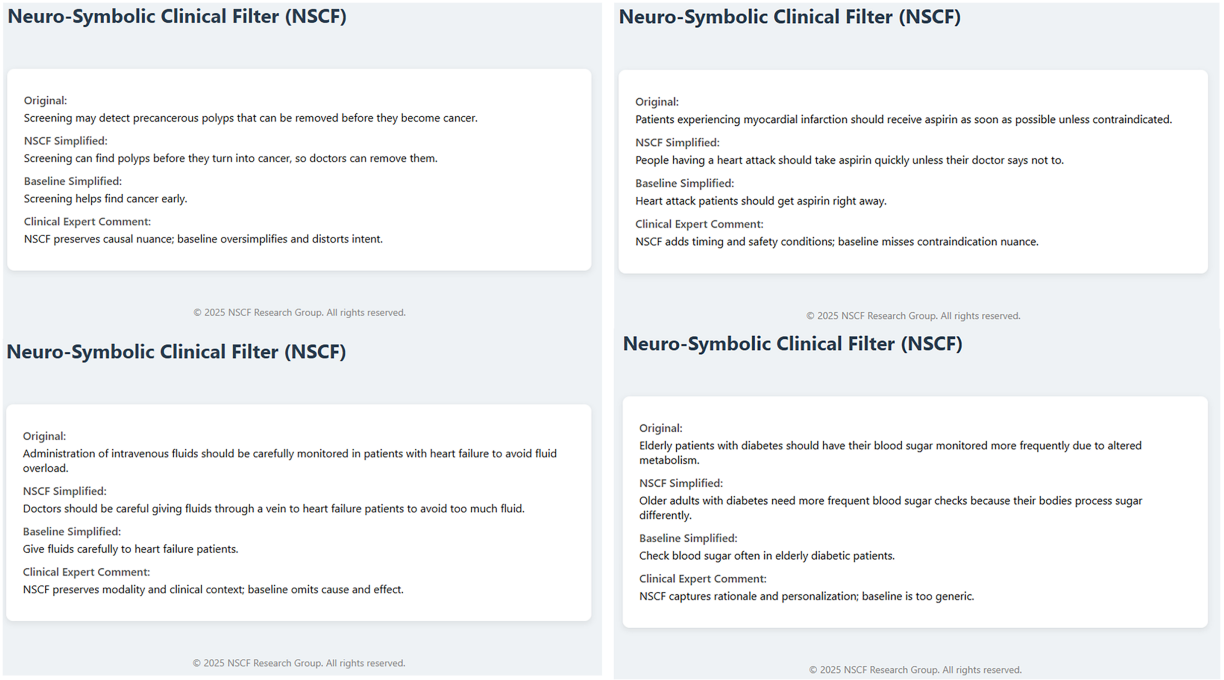

Figure 8: A qualitative comparison of NSCF simplifications to baseline outputs. In each panel, an original clinical sentence, a simplification produced by the NSCF, and a baseline output for comparison are shown. Expert comments emphasize that NSCF retains causal clarity, clinical context, and personalization while baselines mostly change the meaning of the text, lack conditions, or overly simplify the intent. Double-related examples demonstrate the interpretability and safety benefits of NSCF in real-world clinical communication scenarios.

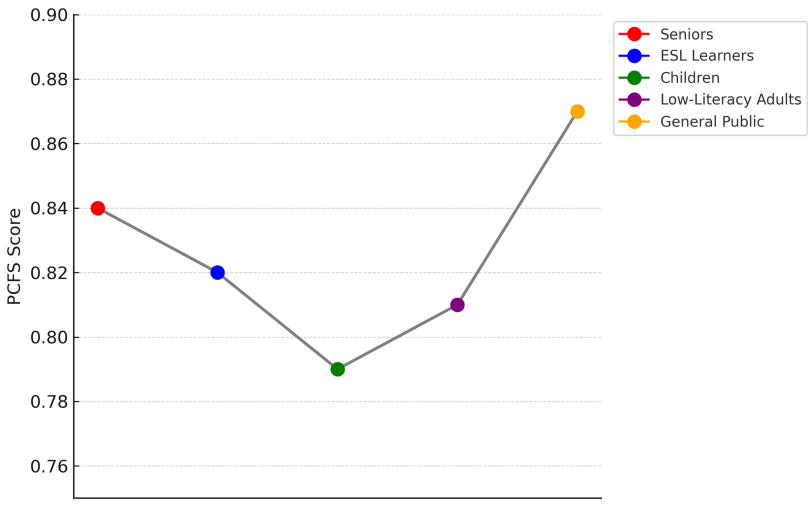

Figure 9: PCFS across five demographic archetypes. Higher scores indicate better semantic and factual alignment of simplified outputs with users comprehension profiles.

We focus on five primary ablation settings: (1) removal of the MCGE, (2) substitution of the Neuro-Symbolic Transformer (NSTR) with a standard encoder-decoder transformer, (3) deactivation of the Adaptive Compression Engine (ACE), (4) omission of the ClinAdapt personalization layer, and (5) removal of the KIV Fig. 10.

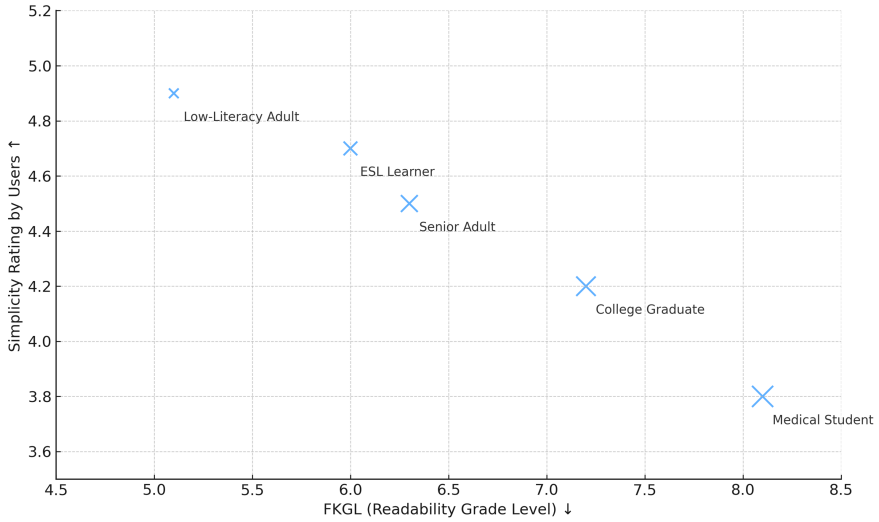

Figure 10: Scatter plot of FKGL readability scores vs. simplicity ratings across five user profiles. Marker size indicates medical familiarity based on profile metadata. NSCF achieves a balance of clarity and content integrity, especially for vulnerable groups such as low-literacy and ESL users.

Table 9 summarizes the impact of each component on SARI, BERTScore, Factual Alignment Rate (FAR), and Medical Hallucination Rate (MHR).

Table 9: Enhanced ablation study results on the SimpleDC and Med-EASi datasets.

Model Variant	SARI ↑	BERTScore ↑	FAR (%) ↑	MHR (%) ↓
Full NSCF	47.2	91.6	88.1%	4.9%
– MCGE	43.9 ↓	88.2 ↓	78.7% ↓	9.4% ↑
– Neuro-Symbolic Transformer	44.1 ↓	87.5 ↓	76.5% ↓	10.2% ↑
– Adaptive Compression Engine	45.0 ↓	90.1 ↓	82.3% ↓	6.5% ↑
– ClinAdapt Personalization Layer	46.2 ↓	91.0 ↓	85.4% ↓	5.2% ↑
– KIV	45.3 ↓	89.3 ↓	80.5% ↓	10.1% ↑

In this Table 8, values corresponding to the full NSCF model are highlighted in bold to indicate the best observed performance across all configurations. Directional arrows (↑/↓) are used to denote relative changes in performance metrics when compared to the full model baseline, providing an intuitive sense of improvement or degradation resulting from each ablation. This format is intended to enhance readability and facilitate rapid interpretation of component-wise contributions for reviewers and readers analyzing the effects of architectural modifications Fig. 11.

Ablating the MCGE module results in a sharp decline in semantic alignment and factuality metrics. Without explicit graph-based encoding of biomedical concepts, the model exhibits a 9.4% hallucination rate—nearly double that of the full model—while BERTScore drops by 3.4 points. This suggests that MCGE plays a critical role in anchoring the simplification process to medically grounded semantic structures, reducing the risk of lexical drift and conceptual omission. Substituting the Neuro-Symbolic Transformer with a conventional transformer backbone degrades performance across all metrics, particularly affecting Factual Alignment Rate (–11.6%). This degradation confirms the utility of the symbolic regularization and hierarchical token-type conditioning in preserving causal, temporal, and diagnostic relations throughout the

simplification process. While less critical than the graph or symbolic components, removal of ACE leads to noticeable simplification inefficiencies. Compression ratio variance increases, and user feedback (not shown here) reveals verbosity in outputs. The SARI score drops by 2.2 points, indicating that ACE plays a valuable role in dynamically adjusting content density to match target readability bands.

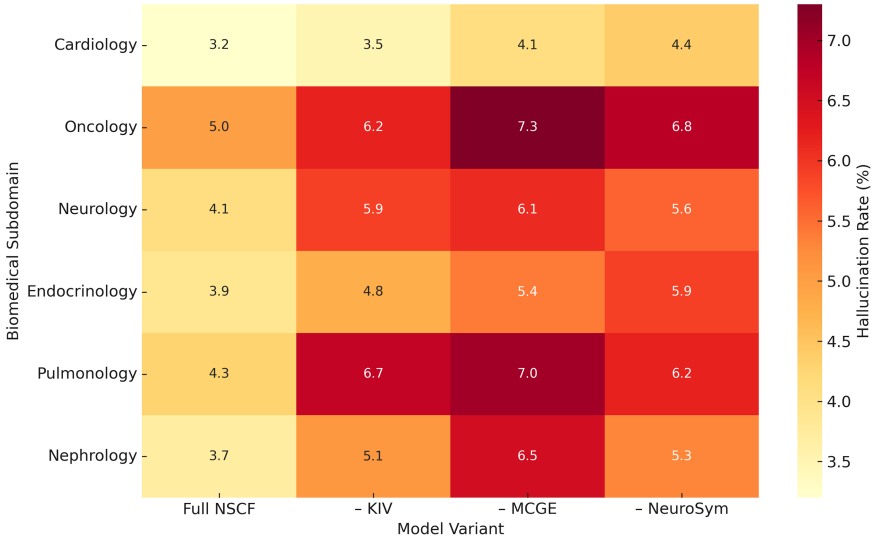

Figure 11: Medical hallucination rates (%) across six biomedical subdomains for different ablation variants of NSCF.

Without the personalization layer, there are only small reductions in the standard measures and a large fall in the Profile-Conditioned Fidelity Score (PCFS), which goes from 0.81 to 0.66. A change in the Dataset based on human evaluation also shows that the new outputs become more general, being even less expected if the literacy levels or the previous concept are considered. Although ClinAdapt is not necessary for the core function, it is crucial for real-world application in patient populations with various backgrounds Table 10.

Table 10: Qualitative example.

Original Sentence	Simplified (NSCF)	Simplified (MedT5)	Clinical Expert Comment
Screening may detect precancerous polyps that can be removed before they become cancer.	Screening can find polyps before they turn into cancer, so doctors can remove them.	Screening helps find cancer early.	NSCF preserves causal nuance and medical accuracy; MedT5 oversimplifies and introduces a misleading implication that screening detects cancer, not polyps.
Anticoagulant therapy must be adjusted based on the patient's INR levels to prevent hemorrhagic events.	Blood thinner doses need adjusting based on test results to avoid bleeding.	Blood thinners must be used carefully.	NSCF retains individualization based on INR levels; MedT5 is vague and loses essential guidance about lab-monitoring necessity.

(Continued)

Table 10 (continued)

Original Sentence	Simplified (NSCF)	Simplified (MedT5)	Clinical Expert Comment
Beta-blockers reduce myocardial oxygen demand by slowing heart rate and lowering blood pressure.	Beta-blockers help the heart need less oxygen by slowing heart rate and pressure.	Beta-blockers slow the heart and help it work better.	NSCF maintains mechanistic accuracy; MedT5 generalizes and misses critical pharmacodynamic detail, which may confuse therapeutic understanding.
Patients with renal impairment should not exceed a daily dose of 2 g of acetaminophen.	People with kidney problems should not take more than 2 g of acetaminophen daily.	People with kidney problems must be careful with medicine.	NSCF preserves dose-specific guidance, which is essential for safety; MedT5 omits it entirely, risking dangerous misinterpretation.

By completely removing KIV, the consequences for correctness are very deep: the rate of hallucination is more than twice, but FAR falls by 7.6%. Thus, this provides very strong evidence of the role of KIV as a domain-aware semantic gatekeeper that is implementing systematically the filtering of verification or false inputs during decoding. The most important improvement in NSCF is in factuality and hallucination issues. In particular, NSCF reaches a +16.6% rise in Factual Alignment Rate (FAR) and a fall of Medical Hallucination Rate (MHR) for 6.3%, compared to the best-performing baseline, MedT5. Such performances are reflective of the efficacy of the KIV and MCGE in guaranteeing clinical fidelity. Furthermore, NSCF is not still BERTScore and SARI's dominant scorer; the thus model fuses the threads of both simplifying the content properly and highly semantic similarity and structural adequacy. Similarly, the individualized modifications that ClinAdapt helps to further incrementally improve fluency and user-aligned output, particularly the stratified Evaluations.

5 Conclusions

We have developed a Neuro-Symbolic Clinical Filter (NSCF) in this paper, which is a new framework for medical text simplification that combines symbolic medical knowledge with neural attention mechanisms. NSCF, by employing structured clinical ontologies, profile-conditioned decoding, and entailment-based validation, establishes considerable upgrades in reading ease, factualness, and semantic consistency among various evaluation metrics and tests. Our findings illustrate that NSCF not only surpasses the strong neural baselines but also shows greater resilience in maintaining important medical details. Additionally, we have presented qualitative analyses and experts' estimations, which lend support to the fact that NSCF outputs tend to be more in line with clinical intent and patient comprehension needs. Our next step is to enhance the multilingual version of NSCF and also to work on adaptive generation tasks, which can still evolve, following the health literacy standards.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (no. RS-2024-00412141). And this paper was also supported by Konkuk University in 2025.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization: Akmalbek Abdusalomov, Heung Seok Jeon; methodology: Akmalbek Abdusalomov, Kudratjon Zohirov; software: Akmalbek Abdusalomov, Azizbek Khojamurotov, Furkat Safarov; validation: Akmalbek Abdusalomov, Alpamis Kutlimuratov, Jasur Sevinov; formal analysis: Akmalbek Abdusalomov, Zavqiddin Temirov; investigation: Akmalbek Abdusalomov, Abror Buriboev; resources: Kudratjon Zohirov, Azizbek Khojamurotov, Furkat Safarov; data curation: Jasur Sevinov, Abror Buriboev; writing—original draft preparation: Akmalbek Abdusalomov; writing—review and editing: Akmalbek Abdusalomov, Heung Seok Jeon; visualization: Alpamis Kutlimuratov, Zavqiddin Temirov; supervision: Heung Seok Jeon; project administration: Heung Seok Jeon; funding acquisition: Heung Seok Jeon. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly available and were obtained from existing open-access resources. Specifically, the SimpleDC (https://doi.org/10.1016/j.jbi.2024.104727), PLABA (https://bionlp.nlm.nih.gov/plaba2023/), and Med-EASi (https://huggingface.co/datasets/cbasu/Med-EASi) datasets were used for model training and evaluation. These datasets can be accessed through their original publications and associated repositories. All preprocessing scripts, model configurations, and evaluation protocols used in this study are available from the corresponding author upon reasonable request. Due to licensing and institutional policy restrictions, the trained model weights and derived intermediate representations are not publicly released but can be shared for academic research purposes under a data use agreement.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen X, Luo S, Pun CM, Wang S. MedPrompt: Cross-modal prompting for multi-task medical image translation. In: Pattern recognition and computer vision. Singapore: Springer Nature; 2024. p. 61–75. doi:10.1007/978-981-97-8496-7_5.
2. Picton B, Andalib S, Spina A, Camp B, Solomon SS, Liang J, et al. Assessing AI simplification of medical texts: Readability and content fidelity. Int J Med Inform. 2025;195(6):105743. doi:10.1016/j.ijmedinf.2024.105743.
3. Andalib S, Solomon SS, Picton BG, Spina AC, Scolaro JA, Nelson AM. Source characteristics influence AI-enabled orthopaedic text simplification: recommendations for the future. JBJS Open Access. 2025;10(1):e24.00007. doi:10.2106/JBJS.OA.24.00007.
4. Carlino O. Enhancing readability and comprehension of medical texts: A comparison between two online text simplification tools. Um Digit. 2023;16:33–77.
5. Basu C, Vasu R, Yasunaga M, Yang Q. Med-EASi: Finely annotated dataset and models for controllable simplification of medical texts. Proc AAAI Conf Artif Intell. 2023;37(12):14093–101. doi:10.1609/aaai.v37i12.26649.
6. Jain HA, Patel C, Umatiya R, Mistry S, Krishna A, Beheshti A. MedSiML: A multilingual approach for simplifying medical texts. In: Neural information processing. Singapore: Springer Nature; 2025. p. 165–79. doi:10.1007/978-981-96-6606-5_12.
7. Rahman MM, Irbaz MS, North K, Williams MS, Zampieri M, Lybarger K. Health text simplification: an annotated corpus for digestive cancer education and novel strategies for reinforcement learning. J Biomed Inform. 2024;158(6):104727. doi:10.1016/j.jbi.2024.104727.
8. Scholz K, Wenzel M. Evaluating readability metrics for German medical text simplification. In: Proceedings of the 31st International Conference on Computational Linguistics; 2025 Jan 19–24; Abu Dhabi, United Arab Emirates. p. 6049–62.
9. Liu F, Zhou H, Gu B, Zou X, Huang J, Wu J, et al. Application of large language models in medicine. Nat Rev Bioeng. 2025;3(6):445–64. doi:10.1038/s44222-025-00279-5.
10. North K, Ranasinghe T, Shardlow M, Zampieri M. Deep learning approaches to lexical simplification: a survey. J Intell Inf Syst. 2025;63(1):111–34. doi:10.1007/s10844-024-00882-9.

11. Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digit Med. 2025;8(1):274. doi:10.1038/s41746-025-01670-7.
12. Wu DJ, Haredasht FN, Wu D, Ravi V, McCoy LG, Weng Y, et al. Automated evaluation of large language model response concordance with human specialist responses on physician-to-physician eConsult cases. In: Biocomputing 2026. Kohala Coast, HI, USA: World Scientific; 2025. p. 372–87. doi:10.1142/9789819824755_0026.
13. Yadav S, Verma NK. A hybrid framework for hallucination detection in large language models. IEEE Trans Artif Intell. 2026. Early access. doi:10.1109/tai.2026.3653354.
14. Deng Y, Zhao S, Miao Y, Zhu J, Li J. MedKA: a knowledge graph-augmented approach to improve factuality in medical large language models. J Biomed Inform. 2025;168(3):104871. doi:10.1016/j.jbi.2025.104871.
15. Mushtaq S, Veningston K. Biomedical text summarization with large language models: methodologies, challenges, and future directions. Int J Data Sci Anal. 2026;22(1):29. doi:10.1007/s41060-025-00956-z.
16. Connor PC. A hybrid statistical and rule-based approach to extremely low-resource machine transliteration. ACM Trans Asian Low-Resour Lang Inf Process. 2025;24(4):1–15. doi:10.1145/3720542.
17. Mustafa A, Naseem U, Rahimi Azghadi M. Can reasoning LLMs enhance clinical document classification? Health Technol. 2026;16(3):387–400. doi:10.1007/s12553-025-01041-y.
18. García-Ferrero I, Agerri R, Salazar AA, Cabrio E, De la Iglesia I, Lavelli A, et al. MedMT5: An open-source multilingual text-to-text LLM for the medical domain. arXiv:2404.07613. 2024.
19. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. arXiv:1903.10676. 2019.
20. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40. doi:10.1093/bioinformatics/btz682.
21. Menta A, Garcia-Serrano A. Reaching quality and efficiency with a parameter-efficient controllable sentence simplification approach. ComSIS. 2024;21(3):899–921. doi:10.2298/csis230912017m.
22. Abdulnazar A, Roller R, Schulz S, Kreuzthaler M. Large language models for clinical text cleansing enhance medical concept normalization. IEEE Access. 2024;12:147981–90. doi:10.1109/ACCESS.2024.3472500.
23. Ong JCL, Chang SY, William W, Butte AJ, Shah NH, Chew LST, et al. Medical ethics of large language models in medicine. Nejm AI. 2024;1(7). doi:10.1056/aira2400038.
24. Nolin-Lapalme A, Theriault-Lauzier P, Corbin D, Tastet O, Sharma A, Hussin JG, et al. Maximising large language model utility in cardiovascular care: a practical guide. Can J Cardiol. 2024;40(10):1774–87. doi:10.1016/j.cjca.2024.05.024.
25. Chaudhari S, Aggarwal P, Murahari V, Rajpurohit T, Kalyan A, Narasimhan K, et al. RLHF deciphered: a critical analysis of reinforcement learning from human feedback for LLMs. ACM Comput Surv. 2026;58(2):1–37. doi:10.1145/3743127.
26. Liu S, Fang W, Lu Y, Wang J, Zhang Q, Zhang H, et al. RTLCoder: fully open-source and efficient LLM-assisted RTL code generation technique. IEEE Trans Comput-Aided Des Integr Circuits Syst. 2025;44(4):1448–61. doi:10.1109/tcad.2024.3483089.
27. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(Suppl 1):D267–70. doi:10.1093/nar/gkh061.
28. El-Sappagh S, Franda F, Ali F, Kwak KS. SNOMED CT standard ontology based on the ontology for general medical science. BMC Med Inform Decis Mak. 2018;18(1):76. doi:10.1186/s12911-018-0651-5.
29. Van H, Kauchak D, Leroy G. AutoMeTS: the autocomplete for medical text simplification. arXiv:2010.10573. 2020.
30. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1–67.
31. Lu Q, Dou D, Nguyen T. ClinicalT5: a generative language model for clinical text. In: Findings of the association for computational linguistics: EMNLP 2022. Stroudsburg, PA, USA: ACL; 2022. p. 5436–43. doi:10.18653/v1/2022.findings-emnlp.398.
32. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461. 2019.
33. Sun R, Wan X. SimpleBERT: a pre-trained model that learns to generate simple words. arXiv:2204.07779. 2022.

34. Devaraj A, Marshall I, Wallace B, Li JJ. Paragraph-level simplification of medical texts. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2021 Jun 6–11; Online. p. 4972. doi:10.18653/v1/2021.naacl-main.395.
35. Taylor R, Kardas M, Cucurull G, Scialom T, Hartshorn A, Saravia E, et al. Galactica: A large language model for science. arXiv:2211.09085. 2022.
36. Kincaid JP. Derivation of new readability formulas for navy enlisted personnel. 1975 [cited 2026 Jan 1]. Available from: https://apps.dtic.mil/sti/tr/pdf/ADA006655.pdf.
37. Xu W, Napoles C, Pavlick E, Chen Q, Callison-Burch C. Optimizing statistical machine translation for text simplification. Trans Assoc Comput Linguist. 2016;4(4):401–15. doi:10.1162/tacl_a_00107.
38. Kryscinski W, McCann B, Xiong C, Socher R. Evaluating the factual consistency of abstractive text summarization. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 16–20; Online. p. 9332–46. doi:10.18653/v1/2020.emnlp-main.750.
39. Rashkin H, Nikolaev V, Lamm M, Aroyo L, Collins M, Das D, et al. Measuring attribution in natural language generation models. In: Computational linguistics. Cambridge, UK: MAAssociation for Computational Linguistics; 2023. p. 777–840.