S-3 Pipeline for Biomedical Text Simplification

By Ansh Vora, Tanish Chaudhari, Sanjeev Hotha & Sheetal Sonawane, Pune Institute of Computer Technology

Notebook for the SimpleText Lab at CLEF 2025

ABSTRACT

Text simplification is an important field in natural language processing that is aimed at improving the readability and understanding of complex text. This makes scientific and technical text more accessible for individuals with limited literacy and comprehension and aids patients in a healthcare setting. We present the S-3 Simplification system – an expert in biomedical text simplification that produces lexically and structurally simplified text, which is semantically fluent, accurate and easy to understand. The system integrated a semantic simplification using T5 models, AMR (Abstract Meaning Representation) -guided structural simplification and BERT-masked modelling with medical thesaurus for context-aware synonym substitution. This approach highlights the effectiveness of a hybridized model for maintaining meaning and fluency while achieving lexical and syntactic simplification.

MOTIVATION

- **Biomedical literature** is growing exponentially.
- Much of it remains **inaccessible** to patients, caregivers, and even practitioners outside niche fields.
- Heavy use of jargon, long sentences, and nested biomedical entities reduces clarity.

- Rule-based methods → risk of oversimplification.
- Purely neural models → may distort critical clinical meaning or introduce hallucinations.

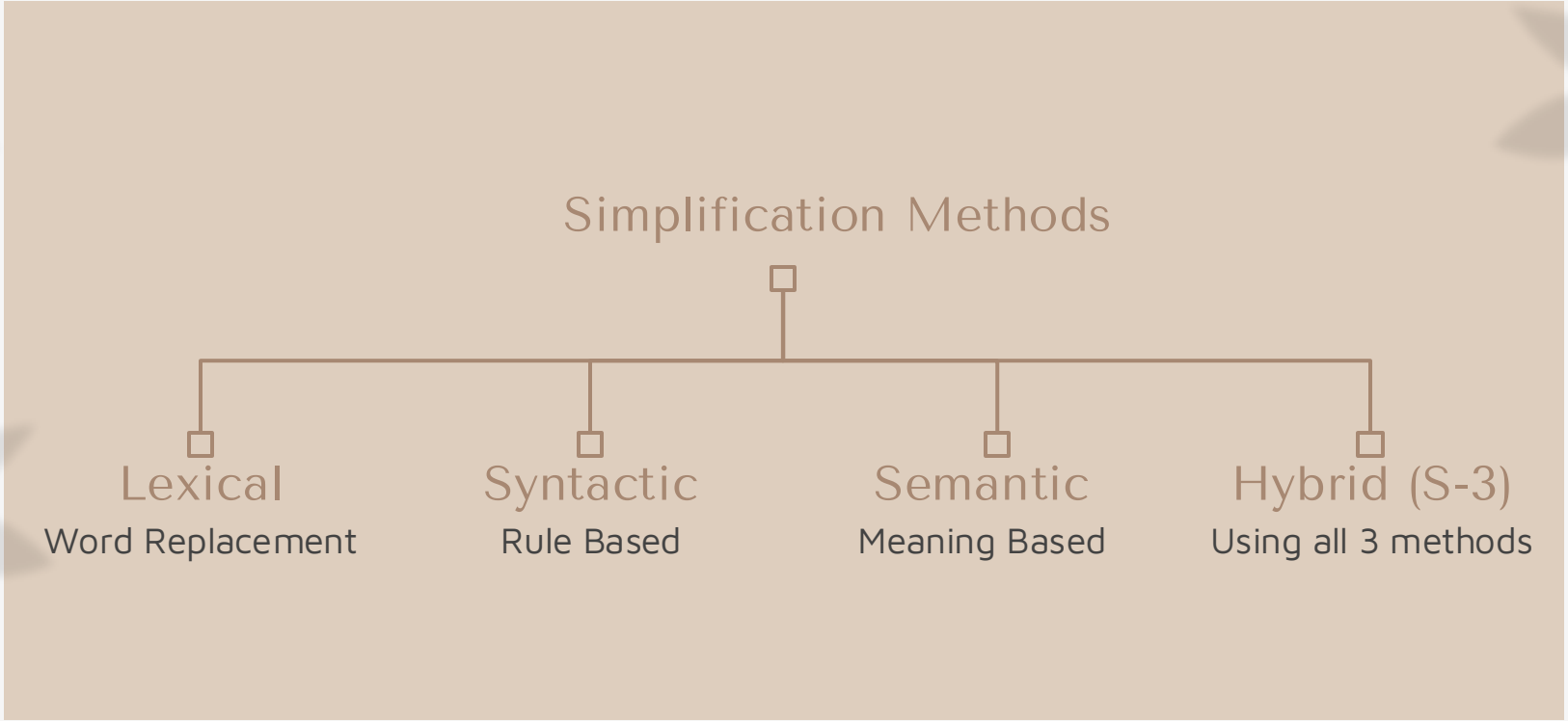
This work addresses the gap by formulating a hybrid biomedical simplification pipeline, **S-3**, which **integrates structural, lexical, and semantic simplification** while preserving key biomedical entities and relations.

The problem is framed around **CLEF 2025 SimpleText Lab Tasks 1.1 and 1.2**, focusing on producing sentences that are simultaneously simpler, semantically faithful, and domain-relevant.

LITERATURE REVIEW

❖ **Classical Lexical Simplification**	Early work relied on rule-based or frequency-based substitution (e.g., SimplePPDB, LexMTR) but lacked contextual nuance for biomedical terms.
❖ **Neural Sentence Simplification**	Seq2Seq and Transformer models (T5, BART) improved fluency but often hallucinate or oversimplify domain-specific content
❖ **AMR for Structural Simplification**	Abstract Meaning Representation has been explored to reorganize complex sentences but rarely applied in biomedical contexts.
❖ **Domain-specific Resources**	WordNet and UMLS provide synonym sets; prior studies used them independently, not as part of a multi-level pipeline.
❖ **Hybrid and Controlled Simplification**	Limited prior research combining structural, lexical, and semantic levels with explicit control tokens for biomedical text.

Approach to Simplification



WHAT SETS US APART?

SUB-LEXICAL

- Focuses on simplifying individual words.
- It replaces complex or rare terms with simpler, more familiar synonyms

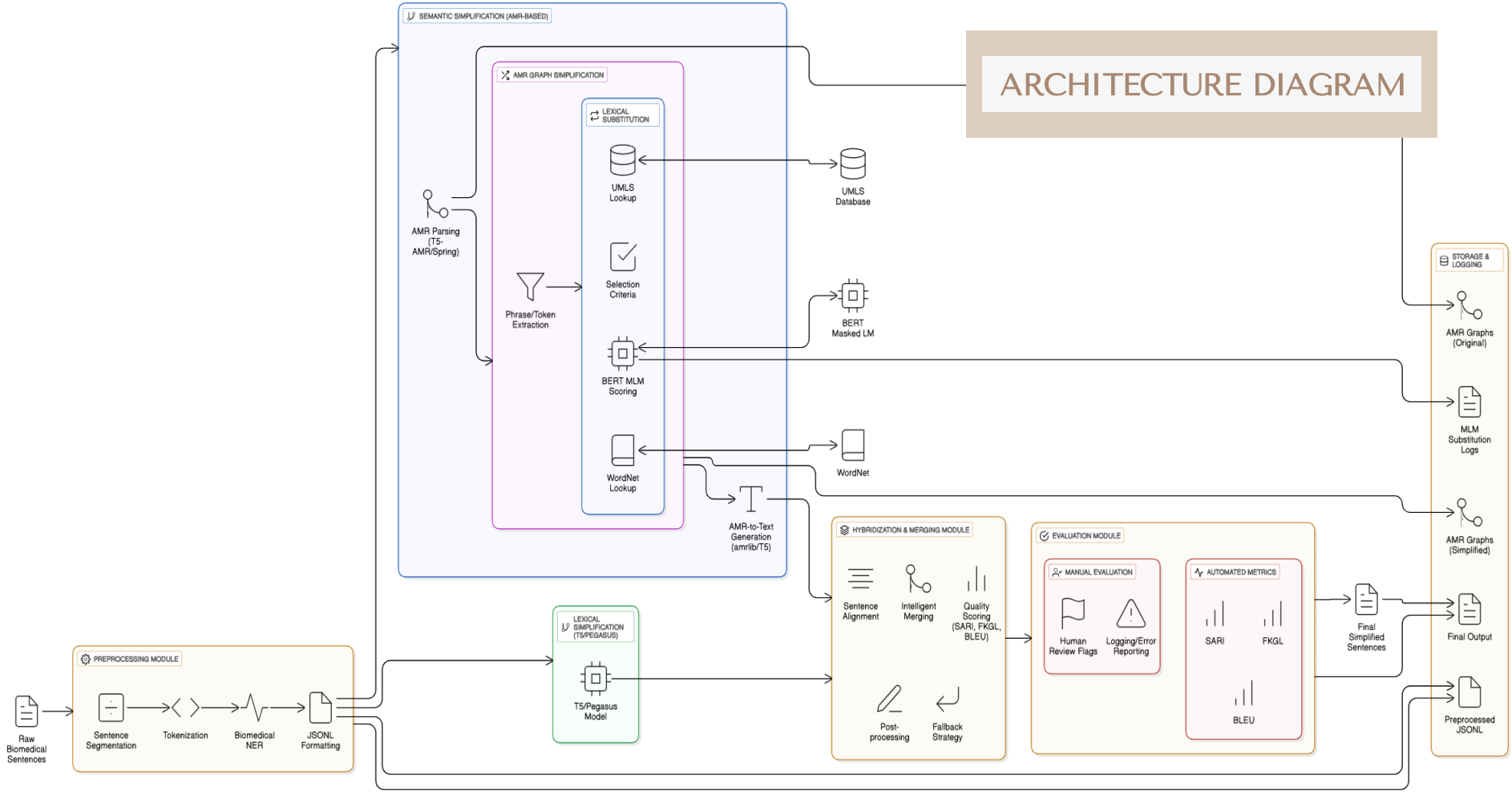
SYNCTACTIC

- Aims to restructure complex sentence syntax
- Splitting long clauses or removing redundant modifiers.
- Relies on parsing to enhance readability.

SEMANTIC

- Targets meaning-level simplification by paraphrasing sentences.
- It involves reduced complexity but maintains semantic fidelity, minimizing information loss.

ARCHITECTURE DIAGRAM



Preliminary Steps

Pipeline Step 1

PRE PROCESSING

- Splits documents into sentences using tokenization techniques.
- Word-level tokenization & POS tagging for lexical replacement.
- SCI BERT tokenizer + SpaCy for domain-specific NLP integration.
- Removes punctuation, citation markers, and performs lemmatization.
- Prepares text for AMR graph generation using AMRlib parser.

Pipeline Step 2

SCI-BERT EMBEDDINGS

- Domain-specific contextual embeddings trained on scientific and biomedical corpora.
- Captures fine-grained biomedical semantics & disambiguates complex terminology.
- Guides lexical substitution candidate ranking via contextual similarity.
- Supports fusion and scoring by evaluating semantic preservation.

Syntactic Simplification

Pipeline Step 3

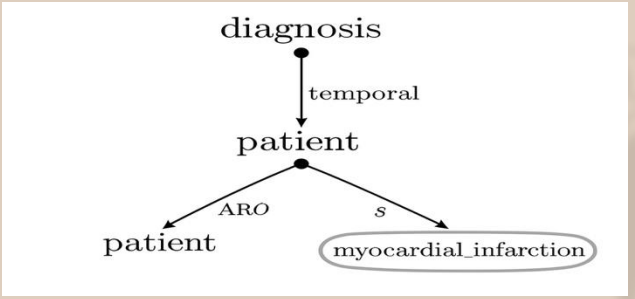
AMR Graph Generation

- Transformer encoder–decoder trained on LDC AMR corpus.
- Graph Representation: Rooted, directed, acyclic graph capturing sentence-level meaning.
- Nodes: Entities, events, attributes, multiword biomedical terms.
- Edges: Semantic roles (agent, patient, modifier) and temporal links.
- Concept Node Identification: Targets semantically dense nodes for simplification.

- Prioritization Criteria:

1. Graph Role – core arguments vs peripheral elements.
2. Domain Specificity – generalizable vs biomedical-bound terms.
3. Lexical Features – TF-IDF weight, document-level importance.

Eg. The patient was diagnosed with myocardial infarction.



Syntactic and Lexical Simplification

Pipeline Step 5A

Lexical Simplification within AMR Graph

•Biomedical sentences often contain few key terms driving complexity (e.g., *myocardial infarction*).
•Applying simplification directly on raw text risks altering non-essential words or grammatical structure.
•AMR graphs isolate core semantic nodes, allowing precise, high-impact simplification.

$$\mathrm{PLL}(x) = -\sum_{t=2}^{T} \log p(x_t \mid x_1, \ldots, x_{t-1})$$

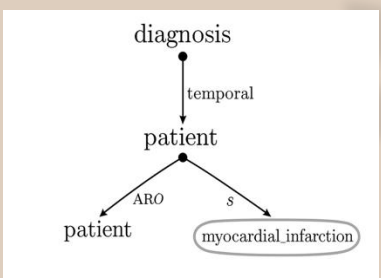
Process:

1.Candidate Generation:

- •Extracts synonyms from WordNet/UMLS based on POS tags (nouns→nouns, verbs→verbs).
- •Filters out overly technical, archaic, or low-frequency terms.
- •Lemmatizes to match AMR conventions.

2.Contextual Evaluation:

- •Uses BERT Masked Language Model (MLM):
- •Replaces target node with [MASK] in the sentence.
- •Each candidate is scored for grammaticality & contextual fit (Softmax + PLL).



Why it works:

•Focuses only on concept nodes → maximizes efficiency.
•Context-aware scoring → prevents meaning distortion.
•Combines frequency & psycholinguistic features → readability improves without clinical loss.

Syntactic and Lexical Simplification

Pipeline Step 5B

Graph Update and Node Substitution

•Simplification must remain **structurally faithful** to the original sentence.
•Updating the graph ensures changes are **semantically anchored**, not superficial replacements.

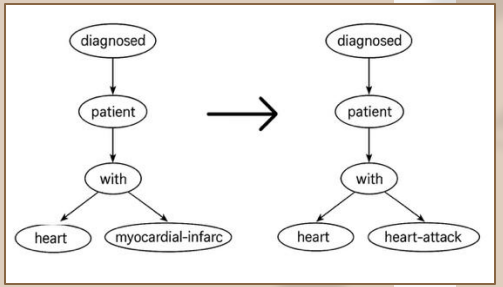
$$\text{BestCandidate} = \arg\max_{c \in \mathcal{C}} \text{PLL}(x_{[p \to c]})$$

Where:
- $\mathcal{C}$ is the set of candidate substitutions (from WordNet/UMLS or any othe rmedical thesaurus or BERT top-k)
- $x_{[p \to c]}$ is the sentence x with phrase p replaced by candidate c

Process:

1.Node Substitution:

•Replace original concept with selected synonym (e.g., (m / myocardial-infarction) → (m / heart-attack)).
•Multiword expressions hyphenated or underscored for AMR compliance.



2.Graph Integrity Preservation:

•Existing semantic edges (:ARGx, :mod, temporal links) remain attached.
•Lemmatization & normalization performed for AMR compatibility.

Why it works:

•Maintains graph connectivity → no loss of argument roles or relations.
•Changes are content-level, not structural → ensures accurate sentence regeneration.
•Prevents errors in downstream sequence-to-sequence generation.

Syntactic and Lexical Simplification

Pipeline Step 5C

Surface Realization

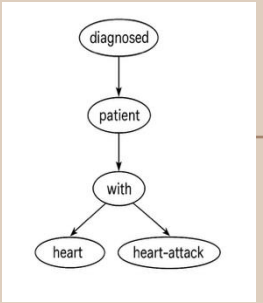
- Converts the **updated semantic graph** back into fluent natural language.
- Ensures simplified terms are **contextually integrated**, not isolated substitutions.

$$\text{Accept}(c) = \begin{cases} \text{True}, & \text{if } \text{PLL}(x_{[p \to c]}) > \text{PLL}(x) \cdot (1 + \delta) \\ \text{False}, & \text{otherwise} \end{cases}$$

Process:

1. **AMR-to-Text Generation:**
Sequence-to-sequence Transformer linearizes graph (PENMAN) → generates sentence.
Inserts function words, adjusts tense and agreement.
2. **Handling Multiword Terms:**
Expands replacements (e.g., "heart-attack" → "heart attack") smoothly into sentence context.
3. **Semantic Fidelity:**
Graph edges preserved → meaning unchanged.
Simplified terms blend naturally with surrounding phrases.

Only nodes are altered → semantic backbone intact.
Surface realization trained on biomedical corpora → **grammaticality retained**.
Allows **fine-grained simplification** without rewriting the entire sentence.



The patient is diagnosed with heart-attack

Semantic Simplification

Pipeline Step 4

Purpose & Role

- Rewrites sentences at a **high level of abstraction** to improve fluency and readability.
- Works **in parallel to lexico-syntactic processing**, not as a replacement.
- Focuses on **whole-sentence meaning**, leaving precise term control to AMR-based phases.

Model & Mechanism

- **Transformer encoder–decoder (PEGASUS, fine-tuned)** for biomedical paraphrasing.
- Captures paraphrastic variation while preserving intent.
- Controlled generation
  - Beam Search (width = 5) for diverse yet focused output
  - Max Length Constraint to avoid verbosity
  - Early stopping for high-probability sequences

Due to insufficient blood flow caused by arterial narrowing, the surgery was postponed to prevent potential complications.

The surgery was delayed because the arteries were too narrow for safe blood flow.

How It Works

- **Input Sentence:** Biomedical text is fed in parallel to Pegasus along with AMR pipeline.
- **Encoding:** Transformer encoder extracts sentence-level meaning and key entities.
- **Paraphrastic Generation:** Decoder rephrases sentence using common, simpler vocabulary.

While Lexico-syntactic simplification involves **precise term substitution and clause restructuring** , semantic simplification involves **contextual sentence rewriting**, improving natural flow and readability while avoiding redundant changes and relies on AMR to protect critical terms.

Hybridization and Merging

Pipeline Step 3

Pegasus:

- Excels at paraphrasing and sentence flow.
- Captures semantic intent but may omit critical details if untuned.

AMR-Lexical:

- Preserves terminology and structure.
- Provides transparent simplification but may sound less natural.
- Aim is to capture both fluency (from Pegasus) and domain precision (from AMR).
- Avoid over-generalization of Pegasus and rigidity of AMR-only outputs.
- In most cases the lexical substitute candidate provides simpler phases but leaves the clause structure a bit rigid, whereas the semantic candidate can rephrase more freely.

1.Input Construction:
Both outputs combined into a single prompt using <L> (Lexical) and <S> (Semantic) tags.
Example: <L> ...lexical tokens... <S> ...semantic tokens...

2.Encoding:
Multi-source T5 encoder maps each token to a vector. Attention layers attend across both streams.

3.Contextualization:
Self-attention learns **which tokens to mix** (phrases from <L>, structure from <S>).

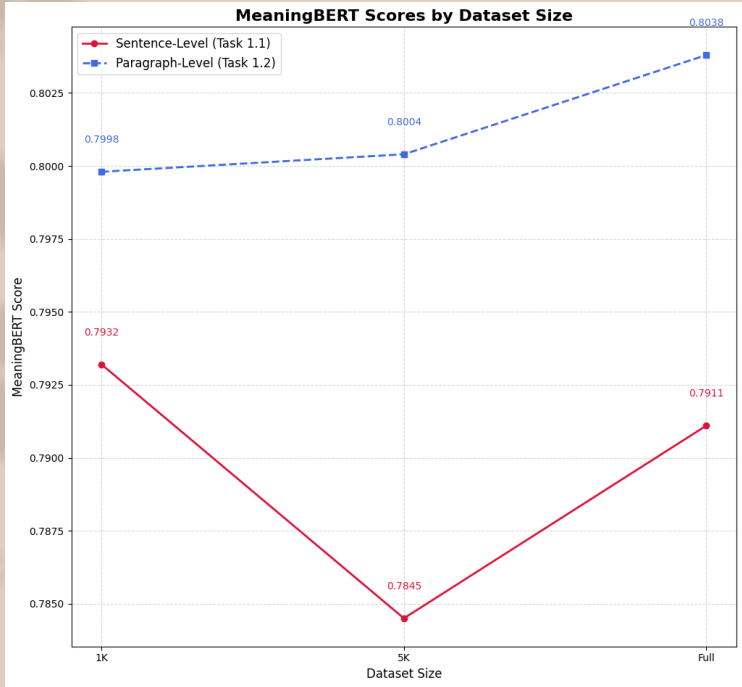
4.Decoding:
Beam search selects top-probability sequence.
Emits final tokens: copies from AMR, reinterpreted phrases from Pegasus, or merged.

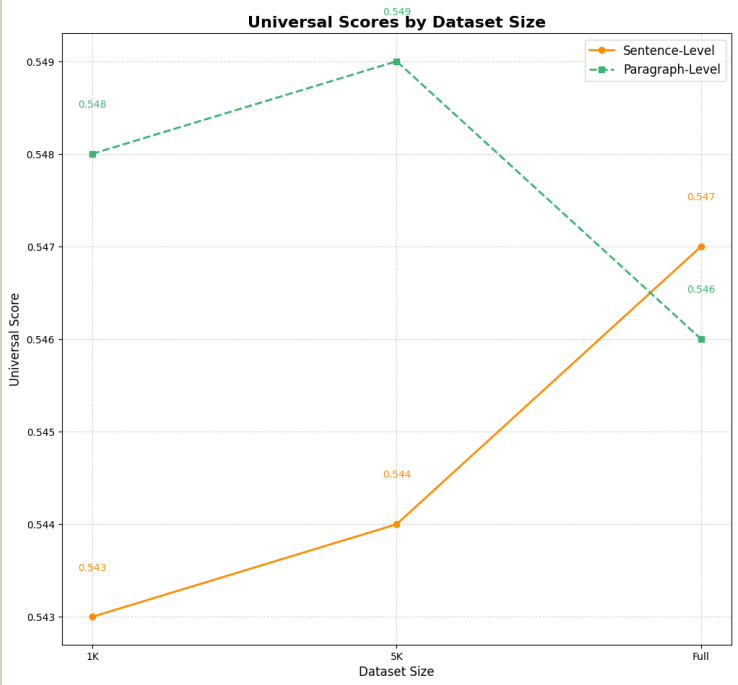
Evaluation Metrics

1.BLEU -- Measures how closely the simplified text matches reference human-written simplifications, based on n-gram overlap.
2.SARI – Evaluates the quality of text simplification by considering how well words are added, deleted, and kept compared to both the original and reference.
3.FKGL – Estimates the U.S. school grade level required to read the text; lower values mean easier readability.
4.ΔFKGL –Shows the change in reading difficulty between the original and simplified text; higher positive values indicate greater simplification.
5.CR –Indicates the proportion of text retained after simplification; lower values mean more text was removed.
6.SLR – Measures how the average sentence length changes after simplification; values near 1 mean minimal change, below 1 indicates shortening.
7.Universal Score- A composite or aggregate score combining multiple metrics to provide a single performance indicator.

$$Score = 0.20 \cdot (1 - BLEU) + 0.20 \cdot \frac{FKGL}{FKGL + 10} + 0.10 \cdot |LR - 0.75| + 0.10 \cdot \frac{PPL}{PPL + 100} + 0.30 \cdot \left(1 - \frac{SARI}{100}\right)$$

Main Metrics





Results for Task 1.1 (Sentence Level Simplification)

Dataset	BLEU	SARI	FKGL	ΔFKGL	CR	SLR	ROUGE-L	Universal Score
1K	0.2782	17.72	10.38	2.58	0.770	0.752	0.534	0.543
5K	0.2696	17.30	10.08	2.76	0.772	0.751	0.531	0.544
Full	0.2806	17.77	10.32	3.14	0.776	0.750	0.544	0.547

FKGL Stability (~10)
•Outputs are simple enough for patients and general readers while retaining biomedical precision.
•This is considered optimal for the biomedical domain, avoiding over-simplification.

ΔFKGL (2.5–3.1)
•Indicates substantial improvement in readability across both tasks.
•Especially effective in paragraph-level datasets, proving the model handles complex, longer sentences well.

SARI Scores (~17)
Ideal for biomedical text: ensures balanced editing without aggressive modification. Falls within the preferred range (16–20), preserving semantic integrity.

Results for Task 1.2 (Paragraph Level Simplification)

Data Size	FKGL	ΔFKGL	SARI	Compression Ratio	ROUGE-L	Universal Score
1K	10.21	2.77	17.43	0.785	0.540	0.548
5K	10.11	2.74	17.21	0.787	0.546	0.549
Full	10.34	3.12	17.68	0.794	0.557	0.546

SLR (~0.75)

- Shows lexical simplification with retention of key terminology & structure. And minimizes risk of factual drift while improving readability.

MeaningBERT (~0.8)

- High scores for paragraph-level simplification, confirming the model's ability to handle large and compound sentences effectively.
- Demonstrates strong semantic preservation with reduced linguistic complexity.

ROUGE-L scores (~0.53–0.55)

- The scores highlight strong content preservation, and Universal Scores (~0.54–0.55) demonstrate consistent, balanced performance across all evaluation metrics.

Performance remains stable from 1K to Full datasets, showing that the model is highly scalable and does not degrade with more data—an essential quality for real-world biomedical applications.

Example

1 Original

Following prolonged administration of corticosteroids, the patient developed osteoporosis, significantly increasing the risk of pathological fractures.

2 Lexico-Syntactic

After long-term use of corticosteroids, the patient developed brittle bones, greatly increasing the risk of abnormal fractures.

3 Semantic

The patient had been on steroids for a long time and their bones became weak, making fractures more likely

4 Hybrid

After long-term steroid use, the patient developed osteoporosis (brittle bones), which raised the risk of serious fractures.

“In this work, the S-3 Pipeline demonstrated excellent performance in biomedical text simplification, effectively combining semantic, structural, and lexical methods. Across metrics such as FKGL, SARI, MeaningBERT, and Universal Score, the system consistently delivered high readability without compromising information content. These results highlight the pipeline’s potential as a reliable tool for improving accessibility of biomedical literature. Future work will focus on enhancing module integration and extending capabilities to multilingual texts.”

—Conclusion

Thank You