NITK SCaLAR Lab at the CLEF 2025 SimpleText Track: Transformer-Based Models for Biomedical Sentence Simplification (Task 1.1)

Notebook for the SimpleText Lab at CLEF 2025

Arya Adwait Dongre[1,†], Ankita Vaadiraaju[1,†] and Anand Kumar Madasamy[1,†]

[1]Department of Information Technology, National Institute of Technology Karnataka Surathkal, Mangalore 575025, India

Abstract

This paper presents the participation of the SCaLAR Lab from the National Institute of Technology Karnataka Surathkal (India) in the CLEF 2025 SimpleText Lab. Biomedical texts are often difficult to understand due to complex vocabulary and sentence structures, which limit access to crucial scientific information for non-expert audiences. Making biomedical literature more accessible, we propose two transformer-based simplification pipelines: one combining BioBERT and BioBART with prompts providing definitions, and another using a fine-tuned GPT-2 Medium model for direct simplification. Our dual approach demonstrates effective reduction of lexical and syntactic complexity while preserving medical accuracy, supporting clearer communication and laying the foundation for future work in multilingual and hybrid simplification systems.

Keywords

Text Simplification, Transformers, Biomedical, BioBERT, BioBART, GPT-2

1. Introduction

The CLEF 2025 SimpleText Track aims to make scientific information more accessible by developing automatic text simplification systems that keep facts correct while reducing linguistic complexity [1]. Scientific literature, especially in the biomedical field, often uses dense vocabulary, complicated sentence structures, and technical jargon that pose major challenges for non-experts like patients, caregivers, and healthcare professionals outside specialized areas [2]. These language barriers hinder the fair spread of important medical knowledge, increasing the risk of misunderstandings or misinformation that can ultimately affect healthcare outcomes [3].

To address these issues, the CLEF 2025 SimpleText Track invites research on natural language processing (NLP) systems that can transform complex biomedical texts into simpler versions while preserving crucial information such as dosages, biomarkers, treatment protocols, and statistical results [4][5]. Automatic text simplification not only improves readability but also supports informed decision-making and promotes inclusivity by bridging the gap between experts and the wider public [6].

The SCaLAR Lab from the National Institute of Technology Karnataka Surathkal (India) participated in CLEF 2025 SimpleText Task 1.1. Our team focused on leveraging large language models (LLMs) for the task, mainly exploring three models: BioBERT, BioBART, and GPT-2.

At the heart of recent advances in text simplification are transformer-based models, especially BERT (Bidirectional Encoder Representations from Transformers) [7], which excels at capturing contextual relationships in sentences using self-attention. However, general-purpose models like BERT often struggle with domain-specific biomedical terminology. This gap is filled by BioBERT, a version of BERT pre-trained on biomedical texts like PubMed abstracts and clinical notes [8], giving it an edge in understanding specialized terms while maintaining their meaning.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[†]These authors contributed equally.

aryadongre.242it010@nitk.edu.in (A. A. Dongre); ankitavaadiraaju.242it004@nitk.edu.in (A. Vaadiraaju); m_anandkumar@nitk.edu.in (A. K. Madasamy)

0009-0006-0947-2938 (A. A. Dongre); 0009-0009-9892-3850 (A. Vaadiraaju); 0000-0003-0310-4510 (A. K. Madasamy)

For generating simplified text, encoder-decoder architectures like BART (Bidirectional and Auto-Regressive Transformers) [9] have become popular in text-to-text tasks. BioBART, which extends BART with biomedical pretraining, can rewrite complex sentences by aligning original tokens with simplified ones through cross-attention, producing clear and accurate simplified texts [10]. Additionally, generative models like GPT-2 [11], even though pre-trained on general data, provide strong baselines for direct simplification when fine-tuned on pairs of complex and simplified biomedical sentences.

In this work, we present two complementary approaches to biomedical text simplification for CLEF 2025 SimpleText Task 1.1. The first combines BioBERT and BioBART with entity recognition from SciSpacy in a multi-step pipeline: first identifying sentences that need simplification, then detecting and defining jargon, and finally using BioBART to generate simplified outputs. The second approach fine-tunes a GPT-2 Medium model on complex-to-simple sentence pairs to enable direct, end-to-end simplification without intermediate steps. Both methods aim to simplify text while preserving key medical concepts, avoiding oversimplification that could compromise accuracy or omit essential details, such as information about control groups in clinical trials [12].

We use the dataset provided in the cited paper [6] for training, validation, and testing. It comprises 11,510 training, 1512 test samples and 1697 validation samples each containing complex-simple sentence pairs annotated with labels such as rephrase, split, merge, ignore, and delete. Columns include complex (original sentence), simple (simplified sentence), label (transformation type), para_id, sent_id, doc_pos, doc_quint, and doc_len, capturing sentence positions and document context. Derived from Cochrane systematic reviews, the dataset covers diverse biomedical topics, providing detailed metadata that enables models to learn context-sensitive simplification strategies to transform technical medical text into lay-friendly language. The rest of this paper is organized as follows: we first review relevant literature on biomedical text simplification and transformer-based NLP models [3][8], then describe our methodology, including data preprocessing, model fine-tuning, and pipeline design. Finally, we present experimental results, analyze and infer, and conclude with insights and future directions for improving biomedical text simplification.

Our experimental results show that the fine-tuned GPT-2 Medium model produced more readable and semantically accurate sentence-level simplifications compared to the BioBERT + BioBART pipeline. The GPT-2 approach effectively preserved essential biomedical concepts while simplifying vocabulary and sentence structure for easier understanding by non-experts, highlighting the promise of direct generative models fine-tuned on biomedical text pairs for accessible and reliable sentence simplification.

2. Literature Review

The CLEF 2025 SimpleText Track tackles the important challenge of making scientific texts easier to understand while keeping them accurate. This section reviews key methods and recent progress in natural language processing (NLP) for simplifying biomedical texts, with a focus on domain-specific models, architectures, and evaluation techniques that help ensure clear and precise communication. Text simplification is a vital NLP task that turns complex, jargon-filled writing into more accessible language, helping a wider range of people—including patients, caregivers, non-native speakers, and those without technical backgrounds—better understand important medical and scientific information.

The development of text simplification methods progressed from rule-based to more sophisticated, data-driven methods, the latter driven by the emergence of machine and deep learning. Conventional methods depended on pre-defined linguistic rules for text simplification, but these methods struggled to scale and generalize over a wide range of text types, languages and domains. In contrast, contemporary deep learning methods, especially using neural networks and large-scale pre-trained models, have shown great success by learning intrinsic simplification patterns from large volumes of data. The latest developments in NLP, specifically with transformer-based models like BERT, GPT, and T5, have created new opportunities for text simplification. These models allow for more context-dependent, flexible simplifications beyond basic lexical or syntactic modifications. They are capable of more advanced transformations and maintaining meaning, while keeping content readable at varying levels[13].

This literature review traces the evolution of deep learning approaches in text simplification, outlining key techniques, challenges, and evaluation strategies. It emphasizes the shift from rule-based systems to neural methods, highlighting recent advances such as large-scale language models, the use of parallel corpora, and the difficulty of preserving semantic meaning. By reviewing current research, it identifies emerging trends and unresolved issues in the development of simplification systems [14].

Transformer-based models like T5 and BioBART are especially effective for simplification. T5 treats it as a sequence-to-sequence task, with the encoder handling complex input and the decoder generating simpler text. BioBART, pretrained on biomedical texts, excels at managing technical content and preserving essential details like dosages and biomarkers through cross-attention mechanisms. Additionally, GPT-2 Medium—a decoder-only model—shows strong fluency when fine-tuned, effectively rephrasing and simplifying text despite lacking explicit source-target alignment. These models illustrate how deep learning facilitates simplification that balances readability, coherence, and factual accuracy [15][16].

The quality of text simplification is typically evaluated using a combination of automated metrics. SARI, a widely adopted metric in simplification research, assesses the quality of additions, deletions, and retentions by comparing system output with human references. It is particularly effective in evaluating whether unnecessary complexity has been reduced without losing essential content. In parallel, readability metrics such as the Flesch-Kincaid Grade Level (FKGL) and Flesch Reading Ease (FKE) are employed to quantify how accessible the output is to a general audience. FKGL estimates the educational level required to comprehend the text, while FKE provides a score indicating overall ease of reading.

Biomedical text simplification introduces specific challenges due to the complexity and density of scientific language. Long-range dependencies, context-sensitive terminology, and the need to preserve domain-critical information—such as drug mechanisms or statistical qualifiers—demand more than surface-level simplification. Transformer-based models, while powerful, are limited by token constraints that may truncate important content. To address these issues, hierarchical encoding strategies and domain-adaptive pretraining are increasingly used. Additionally, hybrid architectures that integrate rule-based components with neural models have shown promise in preserving meaning while enhancing readability [17].

3. Methodology

Biomedical simplification is an important NLP task that aims to make complicated scientific information easier to understand for patients, caregivers, and medical professionals. Biomedical texts often include lengthy sentences and specialized terms that can be hard to grasp without expert knowledge. To tackle this problem, we developed two separate methods using transformer models. The first method uses BioBERT to find complex sentences. It combines this with definitions from SciSpacy and then uses BioBART to simplify the text while keeping its meaning intact. The second method fine-tunes GPT-2 Medium directly on a biomedical simplification dataset to create clearer sentence-level translations from prompts. These two methods use both specific and general models to offer flexible and effective solutions for simplifying biomedical text. The following sections explain the model designs, training methods, and system integration in detail.

3.1. Classification with BioBERT

3.1.1. BioBERT Architecture

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is a domain-specific adaptation of BERT, a transformer-based model developed by Devlin et al. (2018). BERT's architecture is built on the transformer encoder, which uses self-attention mechanisms to capture bidirectional contextual relationships within text. BioBERT extends this by pre-training on large-scale biomedical corpora, including PubMed abstracts (approximately 4.5 billion words) and PMC

full-text articles (approximately 13.5 billion words), in addition to the general-domain corpora used for BERT (e.g., Wikipedia and BooksCorpus). This pre-training equips BioBERT with a deep understanding of biomedical terminology, syntactic patterns, and semantic nuances, making it ideal for tasks in the biomedical domain.

The model architecture consists of several key components. The **input embedding layer** converts input tokens into dense vectors by combining token embeddings, positional embeddings (to capture word order), and segment embeddings (which distinguish different sentences in tasks like question answering). Since our task uses only single-sentence inputs, segment embeddings are uniform. The **transformer encoder layers** in BioBERT (base model) include 12 layers of encoders. Within each layer, the **multi-head self-attention** mechanism computes attention scores across all tokens in the input sequence, allowing the model to weigh the importance of each token relative to others. This enables capturing long-range dependencies and contextual relationships critical for understanding complex biomedical sentences. Following the attention mechanism, **feed-forward neural networks** apply a position-wise fully connected feed-forward network to each token's representation, introducing non-linearity and enhancing feature extraction. **Layer normalization and residual connections** are used to stabilize training by normalizing layer outputs and adding skip connections that preserve information flow across layers. Finally, the **output layer** for classification tasks takes the final hidden state of the special [CLS] token (added to the input sequence) and passes it through a fully connected layer with a softmax activation to produce a probability distribution over the output classes.

In our pipeline, BioBERT is fine-tuned for binary classification to label complex sentences as either rephrase or delete. Sentences labeled rephrase contain valuable content requiring simplification, while delete indicates redundancy or irrelevance. Additional labels—merge, split, none, and ignore—are treated as rephrase, as they imply restructuring rather than removal.

3.1.2. Classification Process

The simplification process starts with **tokenization**, where the complex sentence is broken down using BioBERT's WordPiece tokenizer. This tokenizer splits words into smaller subword units—for example, "cardiovascular" might become "cardi##" and "##ovascular." This approach helps ensure compatibility with BioBERT's vocabulary and allows the model to handle words it hasn't seen before. Special tokens, [CLS] at the beginning (used for classification) and [SEP] at the end (marking sentence boundaries), are added, resulting in a sequence like: [CLS] token1 token2 ... tokenN [SEP].

Next comes **embedding generation**, where the tokenized sequence is transformed into input embeddings by combining token, positional, and segment embeddings. These embeddings are then fed into BioBERT's transformer layers. In the **contextual representation** step, the transformer processes the sequence to generate rich, contextualized representations for each token. The final hidden state of the [CLS] token captures the overall meaning of the entire sentence.

For the **classification** stage, this [CLS] representation is passed through a fully connected layer with two outputs, one for each possible action: "rephrase" or "delete." A softmax activation produces probabilities for these classes, and the one with the highest probability becomes the prediction.

Finally, in the **output decision** step, if the model predicts "delete," the system outputs an empty string, ending processing for that sentence. If it predicts "rephrase," the sentence moves on to the next stage for further simplification.

3.1.3. Fine-Tuning BioBERT

Fine-tuning BioBERT on our labeled dataset involves adapting the pre-trained model to classify biomedical sentences as either "rephrase" or "delete." The process begins with **dataset preparation**, where the training set contains pairs of complex sentences and their corresponding labels, while the validation set is used to monitor performance and adjust hyperparameters. The **loss function** employed is cross-entropy loss, which measures the discrepancy between predicted and true labels and guides the optimization process. For optimization, the **AdamW optimizer** is used with a learning rate typically set

between 2e-5 and 5e-5; a linear learning rate scheduler with warmup is applied to help stabilize training. To reduce the risk of overfitting, **regularization** techniques such as dropout (with a probability of 0.1) within the transformer layers and weight decay are implemented. The model is trained for **3–5 epochs**, using early stopping based on validation performance to further prevent overfitting. Finally, a **batch size** of 16 or 32 is chosen to strike a balance between computational efficiency and stable gradient updates.

Fine-tuning enables BioBERT to learn task-specific patterns, such as identifying sentences with redundant technical details (e.g., methodological specifics irrelevant to a lay audience) versus those containing critical information (e.g., treatment outcomes) that should be rephrased.

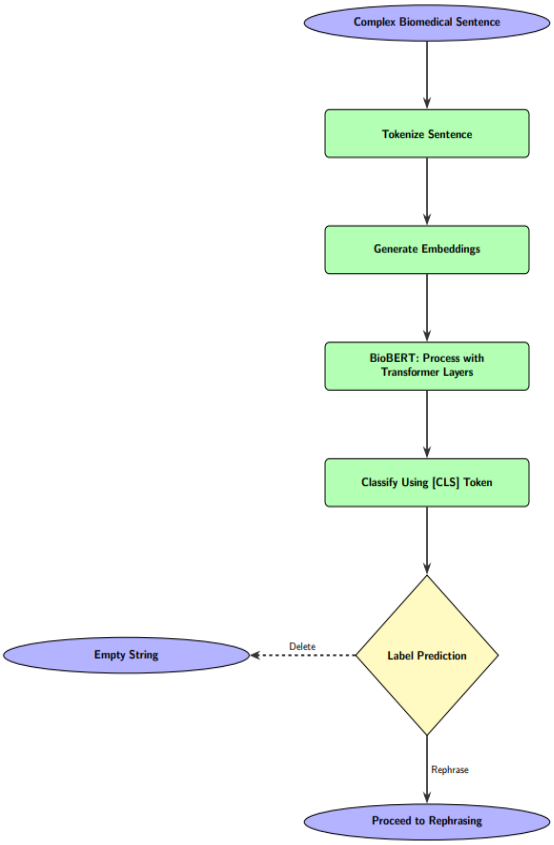

Figure 1: BioBERT Workflow for Classification

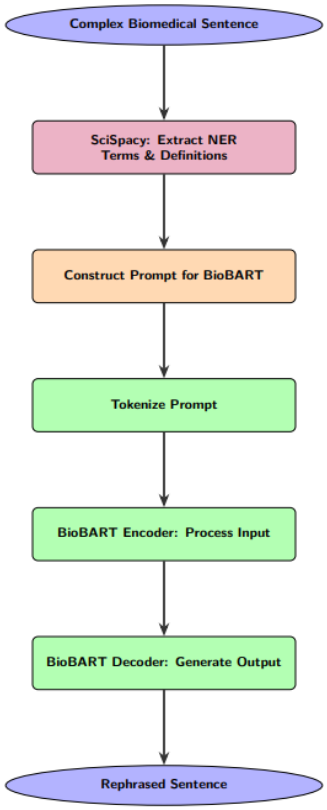

Figure 2: BioBART Workflow for Rephrasing

3.2. Rephrasing with BioBART

3.2.1. BioBART Architecture

BioBART is a biomedical adaptation of BART, a sequence-to-sequence transformer model developed by Lewis et al. (2019). BART combines BERT's bidirectional encoding with GPT's autoregressive decoding, making it ideal for text generation tasks like rephrasing. BioBART is pre-trained on biomedical corpora, ensuring familiarity with domain-specific language.

The model architecture consists of an **encoder** and a **decoder**. The encoder is a bidirectional transformer, similar to BERT, that processes the input sequence—the complex sentence—to generate contextualized representations. It uses 6 layers in the base model, with each layer containing multi-head self-attention and feed-forward networks. The **decoder** is an autoregressive transformer that generates the output sequence—the rephrased sentence—token by token. Like the encoder, the decoder has 6 layers and incorporates cross-attention mechanisms to attend to the encoder's outputs, ensuring that the generated text remains conditioned on the input. **Embedding layers** with shared token embeddings are

used by both the encoder and decoder, complemented by positional embeddings that capture word order; the vocabulary is specifically tailored to biomedical text. Finally, in the **output layer**, the decoder's final hidden states are passed through a linear layer followed by a softmax activation to predict the probability distribution over the vocabulary for each token, enabling text generation.

BioBART's pretraining involves a denoising objective, reconstructing corrupted text (e.g., with masked tokens or shuffled sentences), and equipping it with robust language modeling capabilities. For our task, BioBART is fine-tuned to map complex biomedical sentences to their rephrased counterparts.

3.2.2. Rephrasing Process

The rephrasing process begins with **input tokenization**, where the complex sentence identified by BioBERT as needing simplification is broken down into tokens using BioBART's tokenizer. The resulting sequence is padded or truncated to a consistent length, such as 128 tokens, to ensure uniform input size. In the **encoding** stage, this tokenized sequence is passed through BioBART's encoder, which produces contextualized representations that capture both the meaning and structure of the original sentence. Next, during **decoding**, the rephrased sentence is generated one token at a time, starting with a start-of-sequence token. At each step, the decoder uses cross-attention to incorporate information from the encoder's output—so the generated text stays faithful to the original—and self-attention to maintain coherence by referencing the tokens it has already produced. The next token is chosen by sampling from the probability distribution over the vocabulary, with techniques like beam search used to improve the quality of the rephrased sentence. Finally, in the **output generation** step, the sequence of generated tokens is detokenized to produce the final rephrased sentence.

This output aims to express the original idea in simpler vocabulary and clearer sentence structures, while preserving the intended meaning.

3.2.3. Fine-Tuning BioBART

Fine-tuning BioBART involves training the model on pairs of complex and simplified sentences from the dataset, where the label for each pair is "rephrase."

The process begins with **dataset preparation**, in which the training set provides complex-simple sentence pairs, using the complex sentence as input and the simple sentence as the target, while the validation set supports hyperparameter tuning. The **loss function** used is negative log-likelihood, which measures the difference between the generated and target sequences. For optimization, the **AdamW optimizer** is applied with a learning rate typically between 3e-5 and 1e-4, along with a linear learning rate scheduler to adjust learning dynamics during training. To improve generalization, **regularization** techniques such as dropout (with a rate of 0.1) and label smoothing are included. During training, **teacher forcing** is employed by feeding the ground-truth tokens directly to the decoder to stabilize learning, and during inference, **beam search** with a beam size of 5 is used to enhance the quality of generated sentences. The model is trained for **4–6 epochs**, using early stopping based on validation loss to avoid overfitting. Finally, a **batch size** of 8 or 16 is selected to balance memory constraints with training stability.

Fine-tuning enables BioBART to learn rephrasing strategies, such as replacing technical terms (e.g., "myocardial infarction" with "heart attack"), splitting long sentences, and removing unnecessary details.

3.2.4. Prompt Engineering for BioBART

To enhance BioBART's rephrasing performance, we employ a prompt engineering strategy that provides structured, context-rich input to guide the model. The prompt is designed to simulate the role of a medical assistant, ensuring the rephrased output is clear, accurate, and tailored to non-expert audiences. To investigate how prompt design affects biomedical text simplification, we tested three different prompting strategies with BioBART: (1) a role-only prompt, where the model was told to act only as a medical assistant (e.g., "You are a medical assistant. Simplify: sentence"); (2) a role with terms prompt, which included a list of relevant medical terms along with the role instruction; and (3) a role

with terms and definitions prompt, which also added brief definitions of the key medical terms. The results showed that using both terms and definitions (strategy 3) produced the best performance, with improvements of about 8 to 12 % in BLEU and SARI scores compared to the simpler prompt versions. Human readability assessments confirmed these quantitative improvements. The prompt structure is as follows:

```
You are a medical assistant. Your task is to simplify a complex medical
sentence to make it understandable for non-experts.
Complex sentence: complex sentence
Important terms with definitions:
NER1: Definition
NER2: Definition
...
Simplify the complex sentence using the given information.
```

The prompt used to guide BioBART is carefully structured to ensure the generated simplifications are clear and patient-friendly. It starts with a **role specification**, using the instruction "You are a medical assistant" to set the right context and encourage a professional yet empathetic tone. Next, the **task description** clearly defines the goal with the directive "simplify a complex medical sentence to make it understandable for non-experts," emphasizing the importance of accessibility. The **complex sentence** section simply replaces the placeholder {complex sentence} with the actual sentence identified by BioBERT for rephrasing. In the **important terms with definitions** part, key named entities—such as medical jargon or anatomical terms—are listed along with easy-to-understand definitions, helping BioBART substitute technical language with simpler alternatives. Finally, the **simplification instruction** reiterates the task, reminding BioBART to use the provided definitions to produce a clear, rephrased version of the sentence that preserves the original meaning while making it easier for non-experts to understand.

3.2.5. Named Entity Recognition with SciSpacy

To identify important terms and their definitions, we use SciSpacy, a Python library with pre-trained biomedical NER and entity linking models. First, SciSpacy's NER extracts key entities like diseases or treatments from the complex sentence (e.g., "myocardial infarction," "angioplasty"). Then, in **entity linking**, these entities are connected to UMLS to retrieve standardized definitions, or to simpler lay definitions from a custom dictionary when needed—for example, "myocardial infarction" becomes "a heart attack, where blood flow to the heart is blocked," and "angioplasty" becomes "a procedure to open blocked heart vessels." Finally, the extracted entities and their definitions are inserted into the prompt template, giving BioBART the context it needs for generating clear, simplified rephrasings.

Example Prompt:

```
You are a medical assistant. Your task is to simplify a complex medical
sentence to make it understandable for non-experts.
Complex sentence: Post-myocardial infarction, the patient underwent
angioplasty to restore coronary blood flow.
Important terms with definitions:
Myocardial infarction: A heart attack, where blood flow to the heart is
blocked, causing heart muscle damage.
Angioplasty: A procedure to open blocked blood vessels in the heart using a
balloon or stent.
Simplify the complex sentence using the given information.
```

3.3. Integration of BioBERT and BioBART

The two stages of the pipeline work together in a straightforward and effective way. First, during **input processing**, each complex biomedical sentence is analyzed by BioBERT, which decides whether the sentence should be rephrased or simply removed. At the **decision point**, if BioBERT predicts "delete," the system outputs an empty string right away, skipping any further steps. If it predicts "rephrase," the sentence moves on to BioBART for the **rephrasing** stage, where it's rewritten in simpler, clearer language. Finally, in the **output delivery** step, the pipeline returns either the rephrased sentence or an empty string, depending on the initial prediction. This design makes the most of BioBERT's accuracy in identifying unnecessary content and BioBART's strength in generating high-quality simplifications, while keeping the pipeline flexible so each model can be updated on its own.

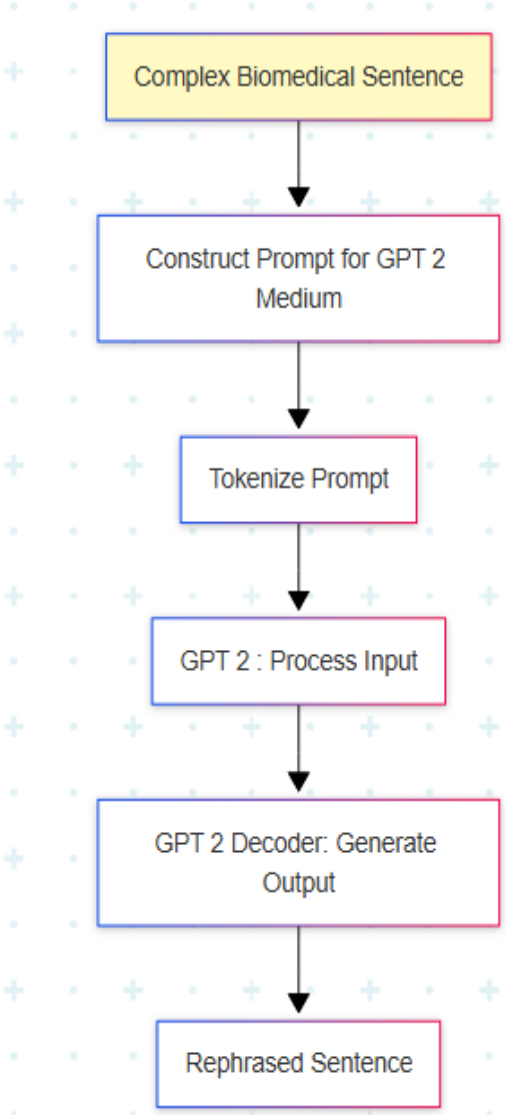

Figure 3: GPT 2 Medium Workflow to simplify

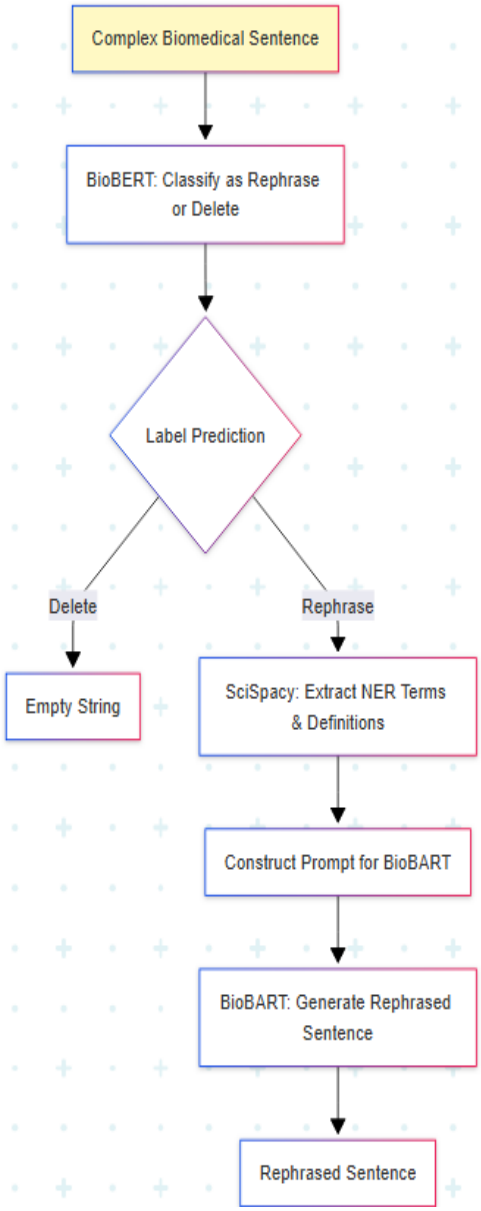

Figure 4: Project Pipeline BioBERT+ BioBART

3.4. Simplification with GPT-2 Medium

3.4.1. GPT-2 Medium Architecture

GPT-2 Medium is a generative transformer-based language model developed by OpenAI, containing 24 transformer layers and 345 million parameters. Unlike encoder-decoder models like BART, GPT-2 Medium operates solely as a decoder in an autoregressive setup. It is pre-trained on a diverse corpus of web text to generate coherent and fluent natural language, making it highly effective for text generation tasks such as simplification when appropriately fine-tuned. Key components of GPT-2 Medium include several architectural elements working together for effective text generation. First, **self-attention layers** are used within each transformer block, employing masked multi-head self-attention so that the model only attends to previous tokens during generation, which preserves causality. Since transformers do not inherently capture word order, **positional embeddings** are learned and added to the input tokens to encode their positions in the sequence. After each attention layer, **feed-forward networks** process the token representations through position-wise fully connected layers, adding non-linearity and improving the model's expressive power. Finally, the **output head** consists of a linear layer followed by a softmax activation, which produces a probability distribution over the entire vocabulary at each time step, enabling the model to generate text one token at a time.

GPT-2 Medium's autoregressive nature makes it ideal for producing natural, fluent simplifications directly from a prompt, without needing a separate encoder.

3.4.2. Simplification Process

GPT-2 Medium simplifies biomedical text through a prompt-based, generative approach. The process begins with **prompt construction**, where each biomedical sentence is preceded by a handcrafted prompt such as "`Simplify the biomedical sentence:\n{text}\nSimplified:`" to clearly cue the model toward the simplification task. During **tokenization**, the combined prompt and sentence are tokenized using GPT-2's tokenizer, with padding or truncation applied to fit a fixed maximum length, typically around 300 tokens. In the **text generation** stage, GPT-2 generates simplified text using controlled sampling strategies: top-k sampling set to 50, top-p (nucleus sampling) at 0.95 for diversity, and a temperature of 0.7 to balance creativity with coherence, while limiting the generation to a maximum of 100 new tokens for conciseness. Finally, during **output extraction**, the generated sequence is decoded and the section following "`Simplified:`" is parsed to produce the final simplified sentence.

This approach enables sentence restructuring, jargon reduction, and verbosity control, making complex biomedical content more accessible.

3.4.3. Fine-Tuning GPT-2 Medium

To tailor the GPT-2 Medium model for biomedical text simplification, we fine-tuned it using a domain-specific dataset composed of sentence-level examples. The aim was to help the model better understand and generate simplified biomedical content while preserving key information.

The training process was carefully designed to help GPT-2 learn to simplify complex biomedical text effectively. Each **input** example began with the prompt "Simplify the biomedical sentence:" followed by a complex medical passage, with the corresponding simplified version serving as the target text. The **learning objective** used Causal Language Modeling (CLM) loss, training the model to predict each token based only on the tokens that came before it. This aligns with GPT-2's autoregressive nature and supports fluent, context-aware generation. For the **training configuration**, the model was trained for three epochs with a learning rate of $5\mathrm{e}{-5}$ and a batch size of one per device. Gradient accumulation over two steps simulated a larger batch size, improving stability on limited hardware. **Mixed-precision** training (`fp16`) was used to speed up computation and reduce memory usage. Progress was logged every 100 steps, and checkpoints were saved at the end of each epoch. For **data collation and tokenization**, `DataCollatorForLanguageModeling` with masked language

modeling disabled (`mlm=False`) ensured compatibility with CLM, while the tokenizer maintained consistent formatting and padding across samples.

This fine-tuning process helped GPT-2 Medium adapt to the linguistic style and complexity of biomedical texts, enhancing its ability to produce simplified content that remains faithful to the original meaning.

3.4.4. Prompt Engineering for GPT-2 Medium

Effective prompt design was key for guiding GPT-2 Medium in simplification tasks. However, prompts were limited to simple role-based instructions. Practical constraints on computational resources and time during experimentation kept us from evaluating more detailed prompting strategies in this setting. The prompt used was:

```
Simplify the biomedical sentence:
<original_text>
Simplified:
<simplified_text>
```

This format encouraged the model to treat simplification as a continuation task, predicting simplified text based on the structure learned during fine-tuning. The inclusion of the explicit keyword `Simplified:` provided a clear delimiter, making the generation task more deterministic and improving the quality of extracted simplifications.

In inference, decoding was performed using nucleus sampling with top-$k = 50$, top-$p = 0.95$, and a temperature of 0.7 to balance fluency and accuracy. The use of the same prompt structure during training and inference ensured consistency and robustness in output generation.

3.5. Hyperparameter Tuning

Hyperparameters for BioBART and GPT-2 Medium were selected by combining limited grid search with standard defaults. Key parameters like learning rate and dropout were initialized from recommended values in prior work and the model documentation. A small set of candidate configurations was tested on a validation set, and final values were chosen based on SARI scores. As extensive tuning was not feasible within practical constraints, most other settings used standard defaults to ensure stable training and reasonable performance.

4. Experimental Results

4.1. Text Simplification Results on Cochraneauto Test Dataset

Task 1.1: Sentence-Level Simplification

Table 1 presents the results of sentence-level simplification for Task 1.1. The table compares the performance of two models, GPT-2 Medium and BioBERT + BioBART, across four evaluation metrics: SARI, BERTScore (F1), FKGL, and FRE. Tested on cochraneauto_sents_test dataset.

Table 1
Sentence-Level Simplification (Task 1.1) Results

Model	SARI	BERTScore (F1)	FKGL	FRE
GPT-2 Medium	**39.7706**	**0.8381**	**9.5214**	**50.05**
BioBERT + BioBART	31.05	0.705	14.3091	27.62

Task 1.2: Paragraph-Level Simplification

Table 2 presents the results of sentence-level simplification for Task 1.2. The table compares the performance of two models, GPT-2 Medium and BioBERT + BioBART, across four evaluation metrics: SARI, BERTScore (F1), FKGL, and FRE. Tested on cochraneauto_para_test dataset.

Table 2
Paragraph-Level Simplification (Task 1.2) Results

Model	SARI	BERTScore (F1)	FKGL	FRE
GPT-2 Medium	32.6018	0.8479	**10.3050**	**47.67**
BioBERT + BioBART	**40.9237**	**0.8862**	19.74	19.40

4.2. Official CLEF 2025 SimpleText Evaluation Results

Task 1.1: Sentence-Level Simplification on Cochrane-auto Abstracts

Table 3 presents the results for CLEF 2025 SimpleText Task 1.1 sentence-level text simplification: Test data on 37 aligned Cochrane-auto abstracts, best five runs per team.

Table 3
CLEF 2025 Task 1.1 Sentence-Level Results on 37 Cochrane-auto Abstracts

Team/Method	*Count*	*SARI*	*BLEU*	*FKGL*	*Compression Ratio*	*Sentence Splits*	*Levenshtein Similarity*	*Exact Copies*	*Additions Proportion*	*Deletions Proportion*	*Lexical Complexity*
Scalar gpt_md_2_1	37	40.95	14.07	18.79	0.62	0.47	0.53	0.00	0.22	0.60	8.68
Scalar BioBart_1	37	33.95	25.69	12.19	0.78	1.00	0.86	0.00	0.01	0.27	8.80
Scalar BioBart	37	33.95	25.69	12.19	0.78	1.00	0.86	0.00	0.01	0.27	8.80

Task 1.2: Document-Level Simplification on Cochrane-auto Abstracts

Table 4 presents the results for CLEF 2025 SimpleText Task 1.2 document-level text simplification: Test data on 37 aligned Cochrane-auto abstracts, best five runs per team.

Table 4
CLEF 2025 Task 1.2 Document-Level Results on 37 Cochrane-auto Abstracts

Team/Method	*Count*	*SARI*	*BLEU*	*FKGL*	*Compression Ratio*	*Sentence Splits*	*Levenshtein Similarity*	*Exact Copies*	*Additions Proportion*	*Deletions Proportion*	*Lexical Complexity*
Scalar gpt_md_2_1	37	34.39	1.01	10.56	0.14	0.19	0.20	0.00	0.03	0.88	8.67

Task 1.1: Sentence-Level Simplification on Plain Language Summaries

Table 5 presents the results for CLEF 2025 SimpleText Task 1.1 sentence-level text simplification: Test data on 217 Plain Language Summaries, best five runs per team.

Table 5
CLEF 2025 Task 1.1 Sentence-Level Results on 217 Plain Language Summaries

Team/Method	Count	SARI	BLEU	FKGL	Compression Ratio	Sentence Splits	Levenshtein Similarity	Exact Copies	Additions Proportion	Deletions Proportion	Lexical Complexity
Scalar gpt_md_2_1	217	38.96	8.25	19.45	0.62	0.43	0.52	0.00	0.23	0.60	8.77
Scalar BioBart	217	30.35	14.26	12.04	0.74	0.99	0.83	0.00	0.01	0.32	8.88
Scalar BioBart_1	217	30.35	14.26	12.04	0.74	0.99	0.83	0.00	0.01	0.32	8.88

Task 1.2: Document-Level Simplification on Plain Language Summaries

Table 6 presents the results for CLEF 2025 SimpleText Task 1.2 document-level text simplification: Test data on 217 Plain Language Summaries, best five runs per team.

Table 6
CLEF 2025 Task 1.2 Document-Level Results on 217 Plain Language Summaries

Team/Method	Count	SARI	BLEU	FKGL	Compression Ratio	Sentence Splits	Levenshtein Similarity	Exact Copies	Additions Proportion	Deletions Proportion	Lexical Complexity
Scalar gpt_md_2_1	217	34.61	0.02	9.26	0.09	0.13	0.13	0.00	0.02	0.93	8.81

4.3. Inference and Analysis

All models were trained exclusively on the provided sentence-level training data to ensure they aligned with the language distribution and specific content of the task. To further test how well these models could generalize, we also evaluated them on paragraph-level texts as an experimental extension, even though they were never explicitly trained on longer passages.

One key finding was GPT-2 Medium's superior readability and SARI scores. As a powerful autoregressive language model pre-trained on a large, diverse corpus, GPT-2 Medium excelled at generating fluent and natural text. When fine-tuned for sentence-level simplification, it consistently produced simpler, clearer sentences while preserving meaning. This was reflected in its strong performance across readability and simplification metrics, indicating effective vocabulary reduction and improved sentence structure without losing important content.

In addition, GPT-2 Medium demonstrated strong semantic preservation. The high semantic similarity scores showed that even after simplifying the surface complexity, it retained the essential meaning

of biomedical content. This highlights GPT-2 Medium's large-scale language understanding, which allowed it to handle the nuances of technical medical text and produce faithful yet easier-to-understand outputs.

However, there were clear limitations in the BioBERT + BioBART pipeline. While this combination leveraged domain-specific knowledge well, it often lagged in readability scores. This can be explained by BioBERT's design: it excels at extracting and understanding biomedical information but wasn't built for generating fluent, readable text. Although BioBART added generative capabilities, the outputs frequently retained dense, technical language, making the simplified sentences still difficult for non-expert readers.

We also observed that high SARI scores don't always indicate effective simplification. When simplifying longer sentences or paragraphs, models like BioBERT + BioBART often preserved much of the original text. Because SARI rewards overlap between the input and output, this led to inflated scores even when the text remained hard to understand. This highlights a mismatch between automatic metrics like SARI and actual readability improvements as perceived by human readers.

A key factor was the domain versus generalization trade-off. The BioBERT + BioBART pipeline tended to prioritize preserving precise biomedical terminology and sentence structure, ensuring accuracy but often at the cost of accessibility. In contrast, GPT-2 Medium applied more aggressive simplification and generalization strategies that improved readability but introduced a potential risk of oversimplification—though this was largely controlled by careful fine-tuning.

We also noted the effects of task granularity. Even though GPT-2 Medium was only trained on sentence-level data, it performed reasonably well on longer sentences, showing some ability to generalize. However, its effectiveness dropped slightly as sentence length and contextual complexity increased, which is expected since longer passages naturally pose greater challenges for simplification.

Finally, the influence of training data was clear. Fine-tuning all models on sentence-level biomedical data helped them learn domain-specific patterns. However, models pre-trained on general language corpora, like GPT-2 Medium, were better able to adapt these patterns into simpler, more readable outputs. Meanwhile, models pre-trained exclusively on biomedical texts, such as BioBERT, tended to simplify more conservatively, limiting their ability to produce easier-to-read rephrasings.

The results demonstrate that GPT-2 Medium, despite not being domain-specialized, is highly effective for biomedical text simplification when properly fine-tuned, especially for improving readability and ease of understanding. Meanwhile, the BioBERT + BioBART approach is valuable when domain fidelity is paramount. Together, these approaches highlight the complementary strengths of general-purpose and domain-specific models, and point toward promising hybrid strategies for future research in multilingual and user-centered scientific communication.

4.4. Implementation Details

All experiments conducted on Kaggle T4 notebooks.

Table 7
Experimental Setup Specifications

Component	Specification
Framework	HuggingFace Transformers v4.47.0
Hardware	2* NVIDIA Tesla T4 (15 GiB GPU max.)
CPU Memory	RAM (29 GiB RAM max.)
Session Disk Size	Disk (57.6 GiB max.)

5. Conclusion and Future Scope

Future research could explore combining specialized biomedical models with general-purpose language models to create even more effective simplification tools and extend these methods to support multiple languages for broader global impact. In this study, we explored two ways to make complex scientific

and biomedical texts easier to understand without losing their core meaning. Both approaches were trained on the provided training data to ensure domain relevance and effectiveness. One approach used BioBERT to identify challenging sentences, enriched them with helpful definitions from SciSpacy, and then simplified them using BioBART. The other approach took a more straightforward path, using GPT-2 Medium to directly rewrite the text in a simplified way. Both methods broke down complicated information into simpler, clearer sentences by removing unnecessary jargon, trimming excess detail, and improving overall readability. Our results showed that each approach successfully made the content easier to grasp while still preserving important medical information. Together, these methods highlight promising steps toward making scientific knowledge more accessible to everyone.

6. Acknowledgments

We thank the organizers of the CLEF 2025 SimpleText track for designing the evaluation framework, providing valuable datasets, and offering timely support throughout the competition. We also appreciate the reviewers' feedback, which helped improve our methods and analysis. Additionally, we acknowledge the contributions of colleagues and mentors who provided insights during model development and experimentation.

Declaration on Generative AI

During the preparation of this work, the authors used *ChatGPT* and *Grammarly* in order to: **Grammar and spelling check** and **Paraphrase and reword**. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Ermakova, H. Azarbonyad, J. Bakker, B. Vendeville, J. Kamps, Clef 2025 simpletext track, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 425–433. doi:https://doi.org/10.1007/978-3-031-88720-8_63.

[2] A. Phatak, D. W. Savage, R. Ohle, J. Smith, V. Mago, Medical text simplification using reinforcement learning (teslea): Deep learning–based text simplification approach, JMIR Med Inform 10 (2022) e38095. URL: https://medinform.jmir.org/2022/11/e38095. doi:10.2196/38095.

[3] W. Coster, D. Kauchak, Simple English Wikipedia: A new text simplification task, in: D. Lin, Y. Matsumoto, R. Mihalcea (Eds.), Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 665–669. URL: https://aclanthology.org/P11-2117/.

[4] R. Chandrasekar, C. Doran, B. Srinivas, Motivations and methods for text simplification, in: COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics, 1996. URL: https://aclanthology.org/C96-2183/.

[5] S. Nisioi, S. Štajner, S. P. Ponzetto, L. P. Dinu, Exploring neural text simplification models, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 85–91. URL: https://aclanthology.org/P17-2014/. doi:10.18653/v1/P17-2014.

[6] J. Bakker, J. Kamps, Cochrane-auto: An aligned dataset for the simplification of biomedical abstracts, in: M. Shardlow, H. Saggion, F. Alva-Manchego, M. Zampieri, K. North, S. Štajner, R. Stodden (Eds.), Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 41–51. URL: https://aclanthology.org/2024.tsar-1.5/. doi:10.18653/v1/2024.tsar-1.5.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2019) 1234–1240. URL: https://doi.org/10.1093/bioinformatics/btz682. doi:10.1093/bioinformatics/btz682.

[9] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. URL: https://arxiv.org/abs/1910.13461. arXiv:1910.13461.

[10] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, S. Yu, Biobart: Pretraining and evaluation of a biomedical generative language model, 2022. URL: https://arxiv.org/abs/2204.03905. arXiv:2204.03905.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[12] Supriyono, A. P. Wibawa, Suyono, F. Kurniawan, A survey of text summarization: Techniques, evaluation and challenges, Natural Language Processing Journal 7 (2024) 100070. URL: https://www.sciencedirect.com/science/article/pii/S2949719124000189. doi:https://doi.org/10.1016/j.nlp.2024.100070.

[13] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, A. Chadha, A systematic survey of prompt engineering in large language models: Techniques and applications, 2025. URL: https://arxiv.org/abs/2402.07927. arXiv:2402.07927.

[14] S. Wubben, A. van den Bosch, E. Krahmer, Sentence simplification by monolingual machine translation, in: H. Li, C.-Y. Lin, M. Osborne, G. G. Lee, J. C. Park (Eds.), Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Jeju Island, Korea, 2012, pp. 1015–1024. URL: https://aclanthology.org/P12-1107/.

[15] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL: https://arxiv.org/abs/1910.10683. arXiv:1910.10683.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. URL: https://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[17] K. Omelianchuk, V. Raheja, O. Skurzhanskyi, Text simplification by tagging, 2021. URL: https://arxiv.org/abs/2103.05070. arXiv:2103.05070.