HAW HAMBURG

BACHELOR THESIS
Nguyen Dinh Hai

Using NLP to Improve Document Accessibility in the Legal Domain

Faculty of Engineering and Computer Science
Department Computer Science

HOCHSCHULE FÜR ANGEWANDTE
WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences

Nguyen Dinh Hai

Using NLP to Improve Document Accessibility in the Legal Domain

Bachelor thesis submitted for examination in Bachelor´s degree
in the study course *Bachelor of Science Informatik Technischer Systeme*
at the Department Computer Science
at the Faculty of Engineering and Computer Science
at University of Applied Science Hamburg

Supervisor: Prof. Dr. Marina Tropmann-Frick
Supervisor: Prof. Dr. Tim Tiedemann

Submitted on: 28. Februar 2025

Nguyen Dinh Hai

Thema der Arbeit

Using NLP to Improve Document Accessibility in the Legal Domain

Stichworte

Rechtsdokumente, Textvereinfachung, Barrierefreiheit, NLP, Juristische Fachsprache, Lesbarkeit, Rechtliche Dokumentenzusammenfassung, ROUGE, BLEU, Textstat, Flesch Reading Ease, SMOG Index, Coleman-Liau Index, Dale-Chall Score, Pegasus, T5, Bart, Summerize, Simplify

Kurzzusammenfassung

Juristische Dokumente enthalten häufig komplexe Satzstrukturen und spezialisierte Terminologie, was ihre Verständlichkeit für Nicht-Experten erschwert. Diese Arbeit untersucht, wie Natural Language Processing (NLP)-Techniken, insbesondere Textvereinfachung und abstraktive Zusammenfassung, die Lesbarkeit und Zugänglichkeit juristischer Texte verbessern können.
In einer Reihe von Experimenten wurden vorgefertigte und feinabgestimmte NLP-Modelle (PEGASUS, BART und T5) evaluiert, der Einfluss der Textvereinfachung auf die Verarbeitung juristischer Dokumente untersucht und die Effektivität verschiedener Verarbeitungskombinationen analysiert. Die Ergebnisse zeigen, dass die Feinabstimmung von PEGASUS auf diverse juristische Datensätze die Zusammenfassungsqualität erheblich verbessert, wobei die Abstractheit und Inhaltsgenauigkeit erhalten bleiben. Allerdings wurde ein Zielkonflikt festgestellt: Während die Vereinfachung die Lesbarkeit verbessert, kann sie durch lexikalische Änderungen zu einem Informationsverlust führen.
Die Reihenfolge der Vereinfachung und Zusammenfassung spielt eine entscheidende Rolle. Die Simplify-then-Summarize-Methode bewahrt juristische Fachbegriffe besser, während die Summarize-then-Simplify-Methode die besten Lesbarkeitswerte erzielt. Bewertungen mit Lesbarkeitsmetriken zeigen, dass die Vereinfachung die sprachliche Komplexität effektiv reduziert und juristische Dokumente zugänglicher macht. Allerdings erfassen standardisierte NLP-Metriken wie ROUGE und BLEU Lesbarkeitsverbesserungen nur unzureichend, da sie primär auf lexikalische Übereinstimmungen und nicht auf inhaltliche Klarheit fokussiert sind.
Diese Forschung trägt zur Entwicklung von KI-gestützten juristischen Anwendungen bei, indem sie zeigt, wie NLP-Techniken helfen können, die Verständlichkeitslücke in juristischen Texten zu überbrücken. Zukünftige Arbeiten sollten Post-Processing-Methoden zur weiteren Verbesserung der Lesbarkeit, menschliche Evaluierungen zur juristischen Genauigkeit sowie Optimierungsstrategien für Vereinfachungsmodelle erforschen, um Zugänglichkeit und juristische Präzision in Einklang zu bringen.

Nguyen Dinh Hai

Title of Thesis

Using NLP to Improve Document Accessibility in the Legal Domain

Keywords

Legal Documents, Text Simplification, Accessibility, NLP, Legal Jargon, Readability, Legal Document Summarization, ROUGE, BLEU, Textstat, Flesch Reading Ease, SMOG Index, Coleman-Liau Index, Dale-Chall Score, Pegasus, T5, Bart, Summerize, Simplify

Abstract

Legal documents often contain complex sentence structures and specialized terminology, making them difficult to understand for non-experts. This thesis explores how Natural Language Processing (NLP) techniques, particularly text simplification and abstractive summarization, can improve the readability and accessibility of legal texts.

Through a series of experiments, we evaluated pretrained and fine-tuned NLP models (PEGASUS, BART, and T5), assessed the impact of text simplification on legal document processing, and analyzed the effectiveness of different processing sequences. The findings show that fine-tuning PEGASUS on diverse legal datasets significantly improves summarization performance while maintaining its abstractiveness and content retention. However, a trade-off was observed: while simplification improves readability, it can lead to a decrease in content retention due to lexical modifications.

The sequence of applying simplification and summarization plays a crucial role. The simplify-then-summarize approach better preserves key legal terms, while the summarize-then-simplify approach achieves higher readability scores. Readability assessments confirm that simplification effectively reduces linguistic complexity, making legal documents more accessible. However, standard NLP evaluation metrics, such as ROUGE and BLEU, may not fully capture improvements in readability, as they primarily focus on lexical overlap rather than conceptual clarity.

This research contributes to legal AI applications by demonstrating how NLP techniques can bridge the gap between legal complexity and accessibility. Future work should explore post-processing methods for readability refinement, human-in-the-loop evaluation for legal accuracy, and optimization of simplification models to balance accessibility with legal precision.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Overview of Legal Documents

Legal documents, particularly terms and conditions (T&Cs), are written in highly formal and complex language, making them difficult for non-expert audiences to comprehend. These documents contain legal jargon, lengthy sentence structures, and domain-specific terms that require expert knowledge to fully understand. Despite their significance in defining rights and obligations, studies indicate that most users do not engage with these documents due to their complexity [34].

Linguistic analyses confirm that legal texts are significantly more complex than standard documents. Table 1.1 provides an example of a typical legal clause, illustrating the long, syntactically complex sentence structures commonly found in legal writing.

Legal Sentence from an SEC Contract Clause

In the event that the Landlord shall deem it necessary or be required by any governmental authority to alter, repair, remove, reconstruct or improve any part of the demised premises or of the building in which the demised premises are located (unless the same result from Tenant's act, neglect, default or mode of operation in which event Tenant shall make all such repairs, alterations and improvements), then the same shall be made by the Landlord with reasonable dispatch, however, such obligation of Tenant shall not extend to maintenance, repairs or replacements necessitated by the intentional wrongdoing or gross negligence of Landlord.

Figure 1.1: Example of a legal sentence from an SEC contract clause [10].

Table 1.1 quantitatively highlights the linguistic differences between legal language and general text, comparing key readability metrics. Legal clauses contain significantly more tokens (129.73 on average) than general English texts (18.16), and their syntactic structure is considerably more complex. Readability scores such as **Flesch Reading Ease** [13], **SMOG Index** [21], **Coleman-Liau Index** [4] further indicate the difficulty of

legal documents compared to general text. These findings emphasize the necessity of text simplification techniques to improve accessibility.

Data Source	# Tokens	Sent. Length	Flesch	SMOG	Coleman-Liau
Legal Clauses (SEC)	129.73	62.52	29.89	35.05	10.79
Simple Wikipedia	18.16	17.98	68.27	8.11	6.83

Table 1.1: Comparison of linguistic complexity between legal clauses and general text (Simple Wikipedia) using readability metrics [10].

1.2 Importance of Enhancing the Understandability of Legal Documents

Legal documents play a vital role across various domains, including business contracts, regulatory policies, and consumer agreements. Their complexity, however, often acts as a barrier to effective comprehension, impacting individuals, organizations, and legal professionals alike. Misinterpretation or lack of understanding of these documents can lead to unintended legal consequences, disputes, and non-compliance issues.

Research has shown that reducing the complexity of legal language can significantly improve comprehension and engagement [14]. By simplifying legal texts, individuals are more likely to understand their rights and obligations, which can foster greater transparency and trust in legal communications. This necessity has fueled growing interest in leveraging Natural Language Processing (NLP) techniques to enhance the accessibility of legal documents. These techniques enable automated simplification and summarization, helping bridge the gap between legal precision and layperson comprehension.

1.3 The Role of NLP in Simplifying and Summarizing Legal Documents

The complexity of legal documents presents a significant challenge for both legal professionals and non-experts. Recent advancements in Natural Language Processing (NLP) have enabled the development of automated solutions that can simplify and summarize these texts, improving accessibility and comprehension. This section explores the role of

NLP in addressing these challenges by leveraging modern techniques for text simplification and summarization.

1.3.1 Text Simplification Using NLP

Legal texts often contain intricate sentence structures, specialized terminology, and lengthy clauses, making them difficult to understand. Traditional rule-based simplification approaches, while effective in specific cases, lack the adaptability required for diverse legal contexts.

For example, the **SIMPATICO** system uses a predefined thesaurus to replace complex words and a rule-based algorithm to select appropriate synonyms [5]. Rule-based methods such as these depend on linguistic databases like WordNet [23] and rely on static rules to determine word difficulty, typically based on frequency or character length. However, this approach does not adapt dynamically to different legal domains and struggles with phrase-level simplifications, making it unsuitable for the complexity of legal texts.

To address these limitations, recent research has proposed learning-based simplification methods, such as the **Unsupervised Simplification of Legal Texts (USLT)** framework [2]. USLT employs a domain-specific legal model, Legal-BERT, to identify and replace complex words while ensuring semantic preservation. Additionally, it incorporates sentence splitting techniques to enhance readability without compromising legal precision.

In the context of legal text simplification, NLP methods typically focus on two main aspects:

- **Lexical Simplification:** This involves identifying and replacing complex words with simpler alternatives using masked language models. The goal is to retain legal meaning while improving comprehensibility.
- **Syntactic Simplification:** Long and convoluted legal sentences are split into shorter, more structured segments to improve readability without losing essential legal nuances.

Unsupervised Learning Methods: Unlike rule-based methods, modern NLP techniques utilize domain-specific corpora to train models capable of automatic simplification without requiring labeled datasets.

1.3.2 Legal Document Summarization Using NLP

Legal professionals frequently deal with extensive documents such as court rulings, contracts, and legislative texts. Extracting relevant information manually is time-consuming and inefficient. NLP-driven summarization techniques, particularly those based on transformer architectures, have demonstrated significant improvements in legal document summarization. Research in *Enhancing Legal Document Summarization Through NLP Models* [12] compares state-of-the-art models, including **T5, Pegasus, and BART**, for generating concise, legally accurate summaries.

To effectively summarize legal texts, NLP-based models employ two primary techniques:

- **Extractive Summarization:** This method identifies and retains the most important sentences from a document while preserving their original wording. It is useful for applications where precise legal terminology must be maintained.
- **Abstractive Summarization:** Unlike extractive methods, abstractive summarization generates new text that conveys the core meaning of the document in a simplified and more readable format. This approach is particularly effective in summarizing complex legal arguments into digestible summaries.

1.3.3 Challenges in Applying NLP to Legal Texts

Despite the progress in NLP, applying these techniques to legal texts comes with unique challenges, as highlighted in [10]:

- **Preserving Legal Meaning:** Any transformation of a legal document must ensure that key legal implications remain intact and legally binding.
- **Lack of Parallel Datasets:** Unlike general-domain text simplification, there are few paired datasets that map complex legal text to simplified versions, making supervised learning approaches difficult to implement.
- **Handling Long Documents:** Many legal documents exceed standard input length limitations of transformer-based models, requiring additional techniques such as hierarchical summarization or document chunking.

- **Evaluation Metrics:** Standard readability metrics may not fully capture the complexity of legal texts. This research employs **Flesch Reading Ease, SMOG Index, Coleman-Liau Index, and Dale-Chall Readability Score** for simplification evaluation and **ROUGE/BLEU** for summarization quality assessment [13, 21, 4, 7].

By distinguishing linguistic complexity (which focuses on inherent textual challenges) from NLP-related challenges (which focus on practical limitations in model training and evaluation), this thesis provides a comprehensive understanding of the barriers to improving legal text accessibility.

1.4 Objectives and Scope of the Thesis

Legal documents, such as contracts, court rulings, and regulatory policies, are often dense and difficult to comprehend for non-expert audiences. This thesis aims to demonstrate how Natural Language Processing (NLP) can enhance the accessibility of legal texts through simplification and summarization. By fine-tuning pre-trained NLP models on domain-specific legal datasets, this research seeks to improve the readability and summarization quality of legal documents while preserving their legal validity.

1.4.1 Research Questions

This thesis investigates the following key research questions:

- **RQ1:** Can NLP-based text simplification and fine-tuned summarization models improve the readability and accessibility of legal documents for non-expert audiences?
  - **Sub-Question:** Do pre-trained summarization models (e.g., PEGASUS, T5, BART) already perform optimally for legal text, or does further fine-tuning on additional legal datasets significantly improve their effectiveness in summarization?
- **RQ2:** Can optimizing the sequence and combination of text simplification and abstractive summarization improve the overall readability and accessibility of legal document?

1.4.2 Thesis Objectives

The main objective of this thesis is to develop and evaluate an NLP-based pipeline that integrates text simplification and abstractive summarization for legal documents. The specific objectives include:

- **Develop an NLP Model:** Design and implement an NLP pipeline utilizing state-of-the-art machine learning algorithms for legal text processing. The focus is on combining simplification and summarization techniques to generate concise, comprehensible legal text.
- **Evaluate Model Effectiveness:** Assess the impact of simplification and summarization on legal text readability using established readability metrics. This includes quantitative evaluations such as text clarity, coherence, and accuracy, as well as qualitative assessments.
- **Promote Legal Accessibility:** Contribute to the broader goal of making legal documents more accessible to non-expert audiences by providing an automated framework for improving legal text comprehension.

1.4.3 Scope of the Thesis

The scope of this research includes:

- **Document Type:** The study will focus on a broad range of legal documents, including contracts, court rulings, regulatory policies, and publicly available legal texts, which are representative of complex legal documents encountered by non-expert users.
- **Language and Jurisdiction:** The research is limited to legal documents written in English to ensure consistency in linguistic and legal structures.
- **Methodology Constraints:** While the project explores cutting-edge NLP techniques, it is constrained by the availability of suitable training datasets and computational resources.

- **Evaluation Metrics:** Readability improvements will be evaluated using **Flesch Reading Ease, SMOG Index, Coleman-Liau Index, and Dale-Chall Readability Score** [13, 21, 4, 7], and summarization quality will be assessed using **ROUGE and BLEU** [17, 26].
- **Use of Transformer-Based Models:** The study will employ transformer-based NLP models such as BART, T5, and Pegasus for legal document processing.

This thesis does not claim to solve all challenges related to legal document accessibility but aims to make significant progress in the application of NLP for simplifying and summarizing legal texts. By establishing clear objectives and a defined scope, this research intends to lay a foundation for future work and provide valuable insights into the use of NLP for improving legal document accessibility.

1.4.4 Definition of Key Terms

In this thesis, the terms "accessibility" and "readability" are used in a precise manner to refer to measurable aspects of legal text processing. These definitions are grounded in prior research on text simplification and legal document processing [10, 2] and are operationalized using widely used linguistic and NLP evaluation metrics.

- **Readability:** Readability refers to how easy or difficult a legal document is to comprehend for a general audience. It is quantitatively assessed using established readability metrics, which have been extensively used in computational linguistics [13, 21, 4, 7]:
  - **Flesch Reading Ease** [13]: A higher score indicates easier readability.
  - **SMOG Index** [21]: Estimates the years of education required to understand a text.
  - **Coleman-Liau Index** [4]: Based on character count per word and sentence length.
  - **Dale-Chall Readability Score** [7]: Considers familiarity with words from a predefined simple word list.

  These metrics were chosen because they have been successfully applied in prior legal text simplification research [10, 2].

- **Accessibility:** Accessibility, in the context of this research, refers to how effectively a legal document conveys essential information to non-expert audiences while minimizing cognitive effort. Since accessibility is not always explicitly defined in legal NLP literature, this thesis operationalizes it using measurable criteria inspired by prior studies on legal text simplification and summarization [10, 2]:
  - **Summarization Effectiveness**: Using **ROUGE** and **BLEU** scores [11] to assess how well the generated summary conveys key legal information.
  - **Cognitive Load Reduction**: The extent to which simplification techniques reduce linguistic complexity, as measured by readability scores.
  - **Usability Analysis**: Whether the restructured text improves ease of reference (e.g., shorter paragraphs, bullet points, clear sectioning).

By defining these key concepts, this thesis ensures that all evaluations are conducted using objective, measurable criteria, enabling a systematic assessment of NLP-based simplification and summarization techniques.

1.5 Thesis Structure and Roadmap

This thesis is organized into multiple chapters, each addressing a key aspect of the research process. Below is an overview of the thesis structure:

- **Chapter 1: Introduction** – Introduces the motivation behind the study, highlighting the complexity of legal documents and the challenges they pose for non-expert readers. It presents an overview of NLP-based solutions, defines the research problem, states the objectives, and outlines the scope of the thesis.
- **Chapter 2: Background and Related Work** – Provides an in-depth review of existing approaches to legal document processing using NLP. It discusses prior work on text simplification and summarization, highlighting the strengths and limitations of different methodologies. The review helps position this thesis within the broader field of legal NLP research.
- **Chapter 3: Methodology** – Describes the methodology used in this research, including:

  - The selection and application of NLP models for simplification and summarization.
  - The preprocessing steps required to prepare legal text data.
  - The design of the experimental setup, including training and evaluation strategies.
  - The fine-tuning strategies, hyperparameter tuning, and computational resource constraints.

- **Chapter 4: Dataset and Model Development** - This chapter details the datasets used for training and evaluation, including their sources, preprocessing steps, and statistical analysis. It also examines dataset characteristics such as word count and sentence complexity to justify their selection for RQ1 and RQ2 experiments. Additionally, the chapter outlines the model selection process and fine-tuning approach for legal text processing.
- **Chapter 5: Experimental Results and Evaluation** – This chapter presents the experimental results and evaluates the performance of the models used for legal document processing. It includes:
  - **Model selection validation:** Verifying if pretraining on BillSum maintains the performance trends observed in prior research for PEGASUS, BART, and T5 models.
  - **Summarization-only evaluation:** Comparing PEGASUS's performance improvements after further fine-tuning using summarization metrics such as ROUGE and BLEU.
  - **Simplification-only evaluation:** Measuring readability improvements through metrics like Flesch Reading Ease, SMOG Index, Coleman-Liau Index, and Dale-Chall Score.
  - **Pipeline evaluation:** Assessing the combined effects of simplification and summarization on overall document accessibility and readability.
  - **Processing sequence experiments:** Investigating whether the sequence of simplification before summarization or vice versa yields better performance.

- **Chapter 6: Discussion and Analysis** – Interprets the findings, discusses their implications, and compares them to prior research. It also identifies the strengths and limitations of the proposed NLP approach.
- **Chapter 7: Conclusion and Future Work** – Summarizes the research contributions and discusses the broader implications of using NLP for legal document processing. It also outlines potential future research directions and acknowledges limitations of the current study.

Each chapter builds upon the previous to provide a comprehensive examination of NLP's role in legal document processing.

2 Motivation and State of the Art in NLP

Legal documents govern fundamental aspects of modern life, from contracts and regulations to privacy policies and court rulings. However, their complexity often acts as a barrier to understanding, leading to unintended legal consequences, disputes, and challenges in compliance. Studies show that a large percentage of the general public struggles to comprehend legal texts due to intricate sentence structures, dense terminology, and domain-specific jargon [10].

Improving the accessibility of legal texts is a growing concern, as discussed in Chapter 1. This thesis focuses on two key aspects of accessibility: **readability**—making legal texts clearer by simplifying structure and vocabulary—and **summarization**—condensing lengthy legal documents while preserving legal intent.

This chapter explores the motivation behind enhancing legal text accessibility and examines state-of-the-art Natural Language Processing (NLP) techniques, particularly modern Transformer-based models, in simplifying and summarizing legal documents.

2.1 Current Challenges in Legal Text Processing

Legal documents are inherently complex, posing significant challenges for both laypeople and legal professionals. While Natural Language Processing (NLP) has made advancements in simplifying and summarizing text, applying these techniques to legal documents introduces unique difficulties. This section categorizes these challenges into two main groups: **linguistic challenges**, which relate to the structure and content of legal texts, and **technical NLP challenges**, which concern the computational and methodological hurdles in processing legal language.

2.1.1 Linguistic Challenges in Legal Text Processing

Legal Complexity

Legal documents employ highly structured language, often embedding multiple clauses within a single sentence. Unlike general prose, legal texts rely on conditional phrasing, regulatory references, and complex syntax, making them difficult for non-experts to understand.

For example, a standard contract clause might read:

> "Neither party shall be liable for any failure or delay in performance under this Agreement (other than for delay in the payment of money due and payable hereunder) to the extent said failures or delays are proximately caused by causes beyond that party's reasonable control and occurring without its fault or negligence, including, without limitation, acts of God, strikes, lockouts, riots, acts of war, epidemics, governmental regulations superimposed after the fact, fire, communication line failures, power failures, earthquakes, or other disasters (each, a 'Force Majeure Event')."

[37]

Such phrasing contains multiple conditions and dependencies, making it difficult to parse. Without explicit restructuring, NLP models may struggle to generate accurate simplifications that preserve the intended legal meaning.

The primary linguistic challenges in legal text include:

- **Syntax Complexity**: Sentences often contain multiple nested clauses and technical jargon.
- **Ambiguity**: Many legal terms have precise but context-dependent meanings.
- **Regulatory Constraints**: Simplifications must retain legally binding language to ensure compliance.

Linguistic Ambiguity

Legal terms often have highly specific meanings that differ based on jurisdiction and context. A single term like "liability" could mean different things in **corporate law, civil law, or criminal law**. This creates ambiguity when trying to simplify or summarize legal text, as models must infer context correctly to avoid misinterpretation.

Readability Barriers

Studies indicate that legal documents frequently score **far below average on readability metrics** compared to general texts [10]. Consider the following readability scores for an average contract:

- **Flesch Reading Ease:** 28 (Highly complex)
- **SMOG Index:** 16.4 (Equivalent to university-level reading)
- **Coleman-Liau Index:** 14.2 (Difficult for non-experts)

This complexity limits accessibility and emphasizes the need for NLP-based solutions that can enhance readability while preserving meaning.

Despite advances in NLP, processing legal texts presents unique challenges, as highlighted in [10]:

- **Preserving Legal Meaning**: NLP-generated summaries and simplifications must ensure that key legal implications remain intact and legally binding.
- **Lack of Parallel Datasets**: Unlike general-domain text simplification, there are few paired datasets that map complex legal text to simplified versions, making supervised learning approaches difficult to implement.
- **Handling Long Documents**: Many legal documents exceed standard input length limitations of Transformer-based models, requiring additional techniques such as hierarchical summarization or document chunking.

- **Evaluation Metrics**: Standard readability metrics may not fully capture the complexity of legal texts. This research employs **Flesch Reading Ease, SMOG Index, Coleman-Liau Index, and Dale-Chall Readability Score** for simplification evaluation and **ROUGE/BLEU** for summarization quality assessment [13, 21, 4, 7].

By distinguishing linguistic complexity (which focuses on inherent textual challenges) from NLP-related challenges (which focus on practical limitations in model training and evaluation), this thesis provides a comprehensive understanding of the barriers to improving legal text accessibility.

2.1.2 Technical Challenges in NLP for Legal Documents

Limited Labeled Data for Specific Legal Domains

While large legal text corpora exist, obtaining labeled datasets for specific branches of law (e.g., tax law, intellectual property law, or regulatory compliance) remains a significant challenge. Many supervised NLP models require high-quality annotations, but labeling legal text demands expert knowledge, making dataset creation costly and time-consuming. This limits the development of specialized NLP models for niche legal domains.. Many legal texts are proprietary, protected under privacy laws, or require specialized domain expertise for annotation. As a result, models trained on general corpora often fail to perform well in the legal domain.

Model Limitations

Many NLP models are designed for **general language understanding** and may not generalize well to **legal terminology and syntax** [3]. Domain-specific adaptations, such as LEGAL-BERT, have been developed to address this issue by pretraining on legal corpora, yet challenges remain.

- **BERT-based models**, including LEGAL-BERT, are limited to a maximum input length of 512 tokens, making them less effective for processing long legal clauses. This constraint often leads to truncation of important information, which impacts downstream legal NLP tasks [3]. Approaches such as Longformer-based adaptations have been introduced to mitigate this issue[1].

- **GPT models** generate fluent text but may introduce factual inconsistencies in legal summaries. Unlike extractive methods, which preserve the original wording, GPT-based abstractive summarization methods sometimes paraphrase or omit critical legal details, raising concerns about reliability [19].
- **T5 and PEGASUS**, while powerful for abstractive summarization, require fine-tuning on legal-specific datasets to maintain accuracy. Their performance degrades when applied to legal texts without domain adaptation, as they lack built-in mechanisms to handle legal terminologies effectively [31].

Without domain-specific adaptation, these models risk generating incorrect or misleading summaries. Continued research in pretraining on larger legal datasets, developing hierarchical architectures, and integrating legal reasoning modules is necessary to enhance the performance of NLP models in the legal domain.

Legal Compliance Issues

Legal documents serve as legally binding agreements, meaning **any errors in simplification or summarization could have real-world legal consequences**. If an NLP model inadvertently alters a contractual obligation or misrepresents a clause, it could lead to disputes or legal liability. This makes accuracy and verifiability critical concerns in legal NLP.

The next section explores existing approachesin detail that address these challenges, focusing on methods for improving legal text simplification and summarization.

2.2 Existing Approaches to Text Simplification and Summarization

The simplification and summarization of legal texts have been widely explored in NLP research, including works such as *Simpatico: A Text Simplification System for Senate and House Bills* [5], *BillSum: A Corpus for Document Summarization of US Legislation* [15], and *Unsupervised Simplification of Legal Texts* [2]. Prior studies have introduced various methodologies to improve legal text accessibility and comprehension. For instance, [5] presents a text simplification system specifically designed for legislative texts, aiming to

make them more accessible to non-expert readers. Similarly, [2] investigates unsupervised simplification techniques to enhance legal document readability.

While rule-based methods have been traditionally used for simplification, modern machine learning approaches have significantly improved summarization quality [15]. The BillSum dataset introduced in [15] facilitates legal document summarization by providing a corpus tailored to the unique challenges of legislative texts, demonstrating the effectiveness of data-driven NLP techniques in summarizing legal documents.

This section discusses existing methods, their effectiveness in addressing challenges outlined in the previous section, and their limitations.

2.2.1 Text Simplification Approaches

Rule-Based Methods

Early text simplification systems relied on predefined rules to transform complex text into more readable forms. These methods included lexical simplification, where difficult words were replaced with simpler synonyms, and syntactic simplification, where long sentences were broken into smaller parts [30]. However, rule-based methods struggle with domain adaptation, often failing to handle domain-specific terminology due to the limited coverage of general-purpose dictionaries [24].

Neural-Based Simplification

Recent advancements have led to neural-based simplification models that use pre-trained Transformers such as **LegalBERT** [3]. LegalBERT, specifically trained on legal texts, helps identify legal terminology and improve contextual simplification. Unlike traditional rule-based methods, neural models leverage large corpora to learn simplification patterns, reducing the need for manual intervention. While these models improve simplification accuracy, they still face challenges in preserving legal intent and precision [24].

2.2.2 Text Summarization Approaches

Extractive Summarization

Extractive summarization selects key sentences directly from the source text while maintaining their original structure. Traditional methods like TextRank [22] apply graph-based ranking to identify the most relevant sentences. While effective, extractive methods often fail to generate coherent and concise summaries suitable for legal applications.

Abstractive Summarization

Abstractive summarization generates new sentences to convey the main ideas of the text in a more concise form. Transformer-based models such as **T5, PEGASUS, and BART** [29, 39, 16] have demonstrated state-of-the-art performance in legal summarization. PEGASUS, for example, is pre-trained using a gap-sentence generation objective, making it particularly effective at summarizing long documents while maintaining fluency. These models address the limitations of extractive methods but require fine-tuning on domain-specific legal datasets to ensure accuracy.

2.2.3 How These Approaches Address Legal NLP Challenges

- **Preserving Legal Meaning:** LegalBERT improves simplification by recognizing legal terminology and maintaining semantic integrity [3].
- **Reducing Readability Barriers:** Neural-based simplification models have shown improvements in readability scores, making legal texts more accessible [10].
- **Improving Summarization Quality:** PEGASUS and BART generate more coherent and fluent summaries compared to extractive methods, hich enhances the accessibility and usability of the summaries for non-expert audiences [32].

While these methods represent significant advancements in legal NLP, further improvements in model adaptation, domain-specific fine-tuning, and evaluation techniques are needed to achieve higher accuracy and reliability. The next section explores the latest developments in NLP, particularly Transformer-based models like BERT, T5, PEGASUS, and LegalBERT, that aim to address these issues and enhance legal text processing.

2.3 State of the Art in NLP for Legal Documents

Advancements in Natural Language Processing (NLP) have significantly transformed legal document processing, particularly in text simplification and summarization. This section presents a comparison of older NLP methods with modern Transformer-based models and highlights domain-specific adaptations such as LegalBERT, LegalT5, and BillSum-trained models.

2.3.1 Comparison of Older and Newer NLP Models in Legal Text Processing

Traditional NLP methods, such as rule-based and statistical models, were widely used in early legal text processing but exhibited limitations in handling linguistic complexity and context sensitivity.

Rule-Based and Statistical Models

Early NLP approaches relied on rule-based systems, where predefined linguistic rules were manually crafted to process text [38]. While these methods were effective for structured tasks such as legal information retrieval, they lacked adaptability and required extensive manual updates. Statistical methods, such as Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs), introduced probabilistic modeling to NLP [20], improving text processing by learning from corpora rather than relying solely on human-crafted rules. However, these approaches still struggled with semantic ambiguity and long-range dependencies in legal text. Semantic ambiguity arises when legal terms have multiple meanings depending on the context, requiring models to accurately interpret their legal intent. Additionally, long-range dependencies in legal documents, such as cross-references to statutes, case law, or contractual clauses, make it difficult for statistical models to maintain coherence across extensive texts. Without deep contextual understanding, these earlier models often misinterpret key legal relationships or fail to capture dependencies that span multiple sections of a document.

Transformer-Based Models for Legal NLP

The emergence of Transformer architectures revolutionized NLP by enabling **context-aware text processing** using self-attention mechanisms [36]. Pre-trained Transformer models such as BERT [8], T5 [29], and PEGASUS [39] demonstrated significant improvements in legal text summarization and simplification tasks compared to statistical methods. Unlike earlier models, Transformers leverage deep contextual embeddings, capturing long-range dependencies crucial for understanding complex legal language.

2.3.2 Domain-Specific Adaptations: LegalBERT, FLawN-T5, and BillSum-Based Models

LegalBERT and FLawN-T5

Recognizing the limitations of general-purpose Transformers on legal documents, domain-adapted models such as **LegalBERT** [3] and **FLawN-T5** [25] were developed. LegalBERT was pre-trained on a corpus of legal documents, enhancing its ability to process legal terminology and complex syntactic structures. Similarly, FLawN-T5 is an adaptation of the T5 framework specifically fine-tuned for legal reasoning tasks. It employs a unique instruction-tuning approach, where the model is trained on a diverse set of legal tasks with natural language instructions. This multi-task learning strategy improves the model's ability to perform various legal tasks, such as summarization, reasoning, and document classification, making it well-suited for handling complex legal texts in multiple legal domains.

BillSum and Legal Summarization Models

One of the most notable contributions to legal text summarization is the BillSum dataset [15], which consists of U.S. Congressional bills and their summaries. Fine-tuning Transformer models such as BillSum-BERT on this dataset has significantly improved legal document summarization accuracy [9]. These models demonstrate how legal domain-specific datasets can enhance Transformer models, leading to more reliable and context-aware legal NLP applications.

2.3.3 Challenges and Future Directions

While modern Transformer-based models represent a significant improvement over earlier NLP techniques, challenges remain in adapting them for various legal domains. Many legal texts require fine-grained contextual understanding, and domain-specific training remains an ongoing area of research. Additionally, balancing readability improvements with legal accuracy remains a critical consideration in legal NLP applications.

The next section will discuss the remaining limitations in current NLP-based legal text processing methods and explore future directions for improvement.

2.4 Current Limitations in NLP-Based Legal Text Processing

Despite significant advancements in NLP, existing models still face critical limitations when applied to legal text processing. These limitations hinder the effectiveness of text simplification and summarization, making legal document accessibility a persistent challenge. This section categorizes these limitations into two main groups: **linguistic challenges** and **technical challenges**.

2.4.1 Linguistic Challenges

Legal Complexity and Structure

Legal texts contain highly structured, hierarchical language, often embedding multiple clauses within a single sentence. Unlike general-domain text, legal documents rely on precise wording, making simplification difficult without altering legal intent. Current NLP models, while effective in general language processing, struggle with maintaining the necessary legal precision in modifications.

Contextual Ambiguity

Legal terminology is highly domain-specific, with many terms carrying distinct meanings in different contexts. For example, the term "liability" may refer to financial obligations in corporate law but to legal responsibility in tort law. Transformer models like BERT and T5 may misinterpret such terms without additional context training on legal corpora [3].

Preservation of Legal Validity

One of the primary concerns with legal text simplification and summarization is ensuring that modifications do not alter the meaning in a way that affects legal validity. Unlike general summarization tasks, where paraphrasing flexibility is allowed, legal summarization must retain key obligations and rights, preventing distortions that could mislead non-expert readers.

2.4.2 Technical Challenges

Domain-Specific Adaptation Needs

Pre-trained Transformer models such as PEGASUS and BART demonstrate strong summarization abilities but often fail in legal domains without additional fine-tuning[12]. Legal text exhibits unique syntactic structures and cross-references, requiring domain-specific training for effective handling. While models like LegalBERT have improved legal text understanding, further adaptations are needed for tasks like legal question answering and argument mining.

Lack of Benchmark Datasets for Simplification

While legal summarization has seen improvements with datasets such as **BillSum** [15], there remains a lack of benchmark datasets specifically designed for legal text simplification. Most simplification datasets cater to general readability improvement (e.g., Simple Wikipedia), which does not reflect the specialized needs of legal text processing. This limitation affects model evaluation and comparative analysis, slowing progress in legal text simplification.

Computational Constraints for Long Legal Documents

Legal documents often exceed the token limits of standard Transformer models, requiring specialized approaches such as hierarchical summarization or chunk-based processing. Without such adaptations, existing models struggle with coherence across long sections, leading to fragmented or inconsistent outputs [39].

While these NLP methods represent major advancements, they still face fundamental obstacles in processing legal text accurately and efficiently. The next section examines how researchers are addressing these limitations through domain-specific training strategies, dataset improvements, and refined evaluation metrics.

3 Methodology

This chapter outlines the methodology used for simplifying and summarizing legal documents using Natural Language Processing (NLP) techniques. The research framework consists of the following stages:

1. **Baseline Evaluation:** Before any preprocessing or model application, the raw legal text is evaluated to establish a baseline for comparison. This step helps measure the effectiveness of the pipeline in improving the readability of legal texts.
2. **Model-Specific Preprocessing:** This step involves preparing legal text for Transformer-based models by applying tokenization, chunking for long documents, and normalization of legal entities.
3. **Model Selection and Training:** Several pre-trained transformer models, including T5, PEGASUS, and LegalBERT, were evaluated based on their performance in legal text processing. Fine-tuned models are compared with pre-trained baselines using summarization metrics (ROUGE and BLEU) and readability metrics (TextStat).
4. **Simplification Development:** A process was established to iteratively develop and refine the simplification step before applying evaluation metrics.
5. **Summarization Evaluation:** The summarization output is assessed post-training using ROUGE, BLEU, and readability metrics such as Flesch Reading Ease and SMOG Index to measure improvements over baseline models.
6. **Simplification Evaluation:** The simplification output is evaluated separately using readability metrics to determine its effectiveness in improving legal text comprehension..

7. **Final Combined Evaluation:** A comparative analysis of the full pipeline's effectiveness is conducted by evaluating how the combined simplification and summarization process improves readability and accessibility compared to the original legal document.

The methodology consists of several interconnected stages, as illustrated in Figure 3.1. Each component is critical to ensuring that legal documents are effectively simplified and summarized while maintaining their legal accuracy.

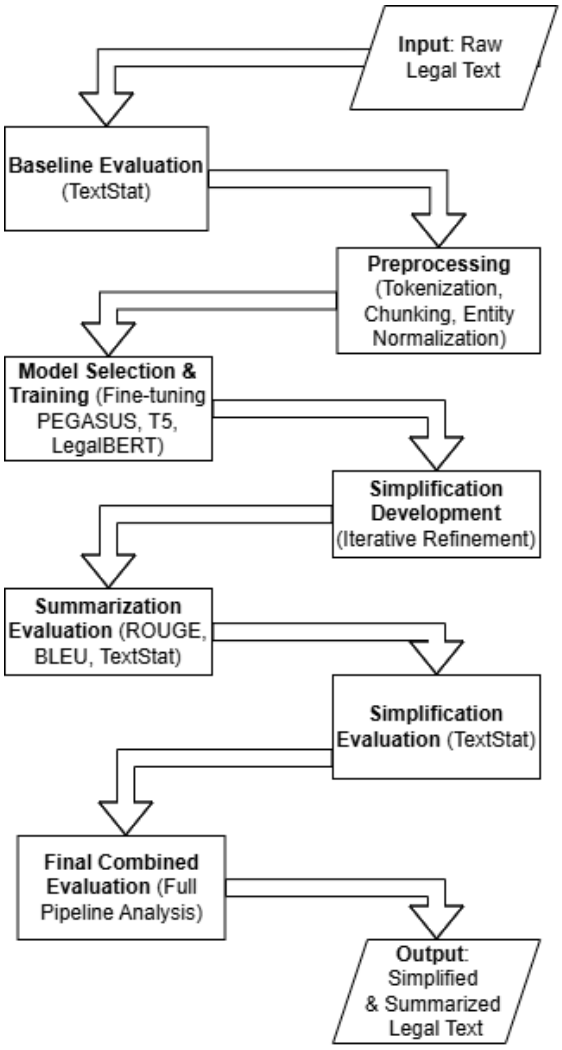

Figure 3.1: NLP Pipeline for Legal Text Processing

In addition to these steps, controlled experiments were conducted to validate the research questions, assessing the impact of fine-tuning on summarization and simplification effectiveness. The experimental results and comparative analysis are discussed in Chapter 5.

3.1 Model-Specific Preprocessing

Preprocessing is a critical step in preparing legal text for transformer-based models, ensuring that raw data is appropriately formatted and optimized for simplification and summarization tasks. The preprocessing pipeline includes tokenization, chunking, and entity normalization, with specific adaptations for legal documents.

While the models selected for fine-tuning were pre-trained on BillSum, additional legal datasets were incorporated to assess whether further fine-tuning on diverse legal texts leads to improved performance. The preprocessing steps described in this section apply specifically to these additional datasets, ensuring consistency in formatting and chunking before model training.

3.1.1 Tokenization and Text Cleaning

Tokenization was applied uniformly across models using the Hugging Face `AutoTokenizer`, ensuring compatibility with transformer-based architectures like T5, PEGASUS, and LegalBERT. This approach allowed for a consistent pipeline while maintaining flexibility for domain-specific tasks.

Minimal text cleaning was performed due to the availability of pre-cleaned online datasets. However, legal structures, such as section headers, citations, and enumerated lists, were retained to preserve the integrity of the legal text. Non-textual elements, such as page numbers, headers, and footers, were excluded to reduce noise.

3.1.2 Chunking for Long Documents

Legal documents often exceed the token length constraints of transformer models, such as 512 tokens for BERT [8] and 512 tokens for T5 [28]. To address this, documents were split into smaller overlapping chunks using a sliding window approach. This strategy ensured

that each chunk retained sufficient context, preventing the loss of critical information at chunk boundaries.

3.1.3 Normalization of Legal Entities

To enhance model understanding, Named Entity Recognition (NER) tools were employed to identify and normalize legal entities, including case citations, laws, and parties. NER is a subtask of Natural Language Processing that automatically detects and categorizes named entities, such as people, organizations, locations, and legal references, within a text. This technique is particularly effective in legal NLP because legal documents frequently contain specialized terms, statutory references, and court names that must be consistently recognized and interpreted.

In this study, NER was used to standardize variations in legal terminology, such as "Supreme Court" and "SCOTUS," ensuring consistency across the dataset. Additionally, custom dictionaries and rule-based mappings were implemented to capture domain-specific legal entities that general-purpose NER models might miss. By applying NER, the system preserves essential legal references, reducing ambiguity and improving the accuracy of both simplification and summarization tasks.

3.1.4 Tools and Libraries

The preprocessing pipeline was implemented using the Hugging Face library, which provided pre-trained tokenizers, NER models, and dataset utilities. These tools streamlined the process of preparing large-scale legal text for model training and evaluation.

3.1.5 Preparing Data for Model Training

After preprocessing, the datasets were combined into a structured `DatasetDict` format using Hugging Face utilities, with separate splits for training, validation, and testing. Each split contained the following features:

- **document:** The raw legal text to be processed.
- **summary:** The reference summary for evaluation.

- **ID:** A unique identifier for tracking data instances.

The dataset consisted of approximately 23,278 training samples, 4,657 validation samples, and 3,105 test samples, providing a robust foundation for model training and evaluation. More on the analysis of the dataset can be found in Chapter 4.

3.2 Model Selection and Training

Selecting the appropriate pre-trained transformer models was a crucial step in developing an effective NLP pipeline for legal document processing. The models used in this study were chosen based on their prior fine-tuning on legal-specific datasets and their demonstrated performance in text simplification and summarization tasks.

3.2.1 Model Selection Process

The models selected for fine-tuning—T5, PEGASUS, and LegalBERT—were pre-trained on diverse corpora, with some, such as PEGASUS and T5, having been further fine-tuned on legal datasets like BillSum. The decision to use BillSum pre-trained models stemmed from their prior exposure to legal domain knowledge, minimizing the amount of additional fine-tuning required.

A prior comparative study on legal document summarization [12] evaluated the performance of T5, PEGASUS, and BART, concluding that T5 excels at capturing semantic subtleties, which is crucial for maintaining the nuanced meanings inherent in legal text. PEGASUS demonstrated superior performance in generating concise abstractive summaries, making it particularly effective for summarizing long legal documents. BART, on the other hand, showed its strength in extractive summarization, preserving structural integrity and key information.

Given these findings, PEGASUS is expected to be the optimal candidate for abstractive summarization of legal texts. This study aims to confirm the applicability of these conclusions when evaluating models fine-tuned on BillSum, with a specific focus on how additional legal datasets may further enhance performance. The comparative evaluation includes:

- Pre-trained models fine-tuned solely on BillSum.

- Models further fine-tuned on additional legal datasets.

An experiment was conducted to validate the effectiveness of PEGASUS for abstractive summarization, comparing T5, PEGASUS, and BART on the fine-tuned dataset. The results, along with the evaluation metrics, are presented in Chapter 5 (Evaluation).

3.2.2 Fine-Tuning Strategy

The fine-tuning process involved training each model on a combined dataset that incorporated legal documents beyond BillSum. Training parameters were optimized to balance text simplification while preserving legal accuracy. Further details on hyperparameters and training configurations are provided in the training section.

3.2.3 Evaluation for Model Selection

To select the best-performing model, initial evaluations were conducted using summarization metrics (ROUGE and BLEU) and readability metrics (Flesch Reading Ease, SMOG Index, Coleman-Liau Index, Dale-Chall Readability Score). These evaluations compared the baseline pre-trained models against their fine-tuned counterparts to measure improvements in readability and summarization accuracy. The results of this comparative analysis are discussed in Chapter 5.

3.3 Experimental Setup

This section describes the experimental setup used to evaluate model performance across multiple aspects of legal text simplification and summarization. The experiments were designed to validate the research questions and assess how different processing techniques impact readability and summarization quality.

3.3.1 Experimental Objectives

The experiments aimed to answer the following key research questions:

- **RQ1:** Can NLP-based text simplification and fine-tuned summarization models improve the readability and accessibility of legal documents for non-expert audiences?
- **RQ2:** Can optimizing the sequence and combination of text simplification and abstractive summarization improve the overall readability and accessibility of legal document?
- **Sub-Question:** Do pre-trained summarization models (e.g., PEGASUS, T5, BART) already perform optimally for legal text, or does further fine-tuning on additional legal datasets significantly improve their effectiveness?

3.3.2 Simplification Process

The simplification process in this study is inspired by the Unsupervised Simplification of Legal Texts (USLT) framework [2]. The architecture is designed to simplify legal texts while preserving their semantic meaning and ensuring readability for non-specialist audiences.

Objectives of Simplification

The primary objective of the simplification process is to enhance the readability of legal texts without compromising their legal integrity. This involves identifying and replacing complex words and phrases while ensuring that the original intent and meaning of the document remain intact. Additionally, the simplification process aims to make legal texts more accessible to a broader audience, including non-specialists, while preserving key domain-specific terminology.

Complex Word Identification

To identify complex words, this study employs a combination of strategies. First, a predefined list of domain-specific legal terms is used to recognize jargon that may be difficult for non-experts. Additionally, word frequency analysis based on corpora such as SUBTLEX is leveraged, applying Zipf scale thresholds to detect low-frequency words that are likely to be complex. Finally, contextual usage and sentence structure analysis help determine whether a word is difficult based on its role within the text [2].

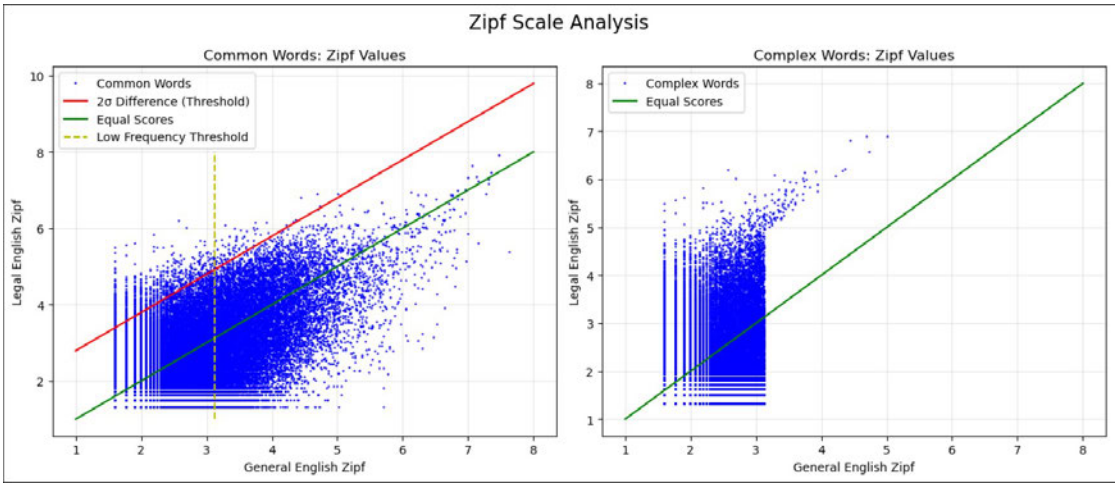

Figure 3.2: Plots of Zipf values of the words common in both SUBTLEX and legal corpora. Adapted from [2].

The graph in Figure 3.3 illustrates how complex words are identified using Zipf scores. The left plot shows common words with higher Zipf values, while the right plot highlights complex words with lower Zipf values. Thresholds are set dynamically based on the dataset to ensure accurate classification.

Substitution Candidate Generation

To generate suitable replacements for complex words, this study leverages **Masked Language Models (MLMs)** such as LegalBERT. These models predict replacement candidates by masking complex words within a sentence and generating contextually appropriate alternatives. The process involves obtaining Top-K Predictions, where the MLM suggests the most probable substitutions based on model confidence and contextual relevance. These ranked candidates are then evaluated to ensure that the best possible simplification is selected without altering the legal meaning of the text.

Weighted Ranking of Substitutions

To select the most suitable replacement for a complex word, this study employs a weighted ranking system that integrates multiple linguistic and statistical factors. Each candidate substitution is evaluated based on five criteria: BERT likelihood, semantic similarity, language model coherence, word frequency, and word length.

The **BERT Likelihood Score** determines how well a replacement word fits within the sentence based on masked language model predictions. To ensure that the meaning of the original term is preserved, the **Cosine Similarity Score** measures the semantic closeness between the original word and its replacement using word embeddings.

Beyond meaning preservation, the **Masked Language Model Score** evaluates how well the candidate maintains coherence within the sentence structure. Additionally, the **Frequency Score**, derived from Zipf scores in SUBTLEX, prioritizes words that are more commonly used in general language. Finally, the **Length Score** introduces a slight penalty for longer words to encourage the use of more concise substitutions.

Each of these factors is weighted according to prior findings from the *Unsupervised Simplification of Legal Texts* paper [2]. The ranking prioritizes words that fit well within the sentence context, closely match the original meaning, and improve readability while maintaining legal accuracy. The assigned weights favor contextual fit the most, followed by semantic similarity, coherence, word frequency, and length penalties.

This ranking approach ensures that word substitutions align with legal text requirements by reducing complexity while preserving essential meaning.

Sentence Reconstruction

After complex words are replaced, the sentence structure is carefully assessed to ensure clarity and coherence. The chosen substitutions are integrated while maintaining the original sentence framework, preventing any distortion of meaning. Additionally, formatting elements such as punctuation and spacing are preserved to uphold the structural integrity of the document. This step ensures that the simplified text remains legally accurate and visually consistent with the original document.

Example Transformation

To illustrate the simplification process, consider the following legal sentence:

Original Sentence:

> "We cannot give you away because for us, you are indispensable to prosecute the plaintiff who committed the crime."

Simplified Sentence:

> "We cannot give you away. For us, you are entitled to be the person who committed the crime."

Output Module

The final stage of the simplification process ensures that the generated outputs are both transparent and structured for further analysis. Each simplified sentence is displayed alongside its original counterpart, allowing for direct comparison and validation of modifications. Additionally, the processed text is exported in structured formats such as CSV and Excel, facilitating its integration into downstream applications and further analysis.

Workflow Description

The complete workflow is as follows:

1. **Complex Word Identification**: Flags words that require simplification.
2. **Candidate Generation and Ranking**: Generates and ranks substitution candidates.
3. **Sentence Reconstruction**: Simplifies and reformulates flagged sentences.
4. **Quality Control and Output**: Validates results and exports simplified text.

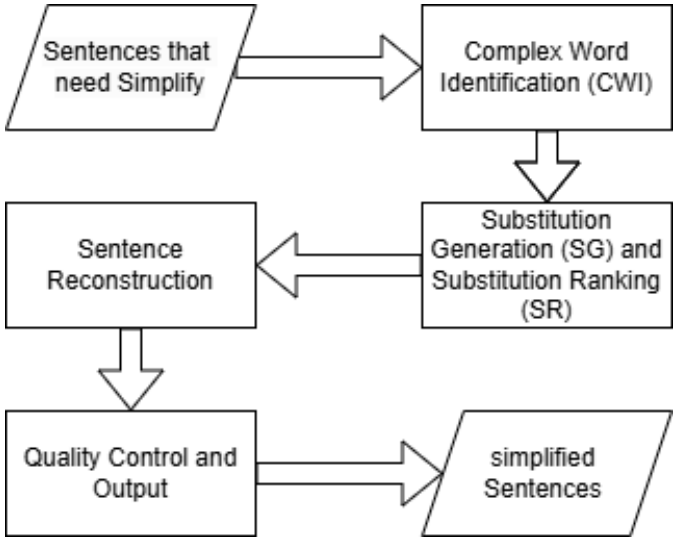

Figure 3.3: Zipf Scale Analysis: Identification of Common vs. Complex Words in Legal Texts

3.3.3 Experiments for RQ1: Evaluating the Impact of Fine-Tuning

To assess whether fine-tuning improves readability and accessibility, this study compares two model configurations. The first consists of baseline pre-trained models - T5, PEGASUS and BART - fine-tuned exclusively on the BillSum dataset. The second extends this process by further fine-tuning these models on additional domain-specific legal datasets, incorporating a broader range of legal texts mentioned on the Chapter 4 dataset, with diversity in model size in terms of words and sentence count.

The primary goal of this experiment is to determine whether exposure to diverse legal texts improves readability and summarisation performance beyond that provided by BillSum training alone. To measure the effectiveness of each approach, the models were evaluated on two key aspects: summarisation quality and readability. Summarisation performance was assessed using ROUGE-1, ROUGE-2, ROUGE-L and BLEU scores, while readability improvements were quantified using established metrics such as Flesch Reading Ease, SMOG Index, Coleman-Liau Index and the Dale-Chall Readability Score.

3.3.4 Experiments for RQ2: Optimizing Simplification and Summarization Order

The sequence in which text simplification and summarization are applied could significantly influence the readability and effectiveness of legal document processing. While both steps serve to improve accessibility, the question remains: **Does simplifying legal text before summarization yield better results than summarizing first and then simplifying the condensed text?**

To investigate this, two processing orders were tested. The first approach applied **simplification before summarization**, where legal documents were initially simplified before undergoing abstractive summarization using a fine-tuned PEGASUS model. This strategy aimed to reduce complexity at an early stage, allowing the summarization model to operate on a more accessible version of the text, potentially leading to more coherent and digestible summaries.

The second approach reversed this sequence by applying **summarization before simplification**, where the document was first summarized and then the resulting summary underwent a simplification process. This method tested whether extracting key legal information before simplification would better preserve essential legal details, ensuring that the most crucial aspects remained intact while still improving readability.

To determine which processing order is more effective, both strategies were evaluated using **readability metrics** (Flesch Reading Ease, SMOG Index, Coleman-Liau Index, and Dale-Chall Readability Score) and **summarization metrics** (ROUGE and BLEU). The results provide insights into whether the order of operations impacts document accessibility and overall comprehension.

3.4 Computational Setup

The training and evaluation experiments were conducted on a high-performance computing server with the following specifications:

- **GPU:** NVIDIA 3090 (25GB VRAM)
- **RAM:** 135GB
- **Training Framework:** Hugging Face Trainer API

- **Experiment Tracking:** Weights & Biases (WandB)

Training Stability and Monitoring

To ensure robust model performance and prevent overfitting, multiple regularization techniques were applied during training. **Dropout regularization** was implemented to deactivate random neurons, improving generalization by reducing reliance on specific features [33]. Additionally, **weight decay** penalized overly large weights, enhancing model stability and preventing excessive complexity in learned representations [18].

Beyond these regularization methods, **early stopping** was used to monitor validation loss and halt training when improvements plateaued, reducing the risk of overfitting [27]. Model checkpoints were saved periodically to ensure that the best-performing version was retained. To track progress and visualize training performance across epochs, Weights & Biases (WandB) was used for logging and experiment management.

3.5 Fine-Tuning Strategy for PEGASUS

Based on the results from **Experiment 1** (see Section 5.1), PEGASUS was validated as the most effective model for summarization tasks, outperforming T5 and BART on ROUGE and BLEU scores. Subsequently, PEGASUS was further fine-tuned on the combined legal dataset to improve its performance on domain-specific legal texts.

The hyperparameters for fine-tuning were replicated from the PEGASUS pre-training study [39], as these settings yielded strong results on BillSum. The key hyperparameters are listed in Table 3.1.

Table 3.1: Hyperparameters used for PEGASUS fine-tuning

Hyperparameter	**Value**
Learning rate	2e-4
Label smoothing	0.1
Number of steps	100k
Beam size	8
Max input tokens	1024
Max target tokens	256

3.6 Evaluation Metrics

This chapter presents the evaluation framework used to assess the effectiveness of the summarization and simplification models developed in this thesis. The evaluation focuses on two key aspects: statistical metrics, which measure how closely the generated summaries align with expert references, and readability metrics, which evaluate how easily non-experts can comprehend the simplified texts. By combining these approaches, the analysis provides a comprehensive assessment of both content fidelity and accessibility.

3.6.1 Summarization Evaluation Metrics

3.6.2 ROUGE Score

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric [17] is widely used to measure summarization quality by comparing model-generated summaries to human-written references through n-gram overlap.

The ROUGE-N score is calculated as:

$$\text{ROUGE-N} = \frac{\sum_{s \in \text{ReferenceSummaries}} \sum_{gram_n \in s} \text{Count}_{match}(gram_n)}{\sum_{s \in \text{ReferenceSummaries}} \sum_{gram_n \in s} \text{Count}(gram_n)} \tag{3.1}$$

where $gram_n$ represents the n-gram, and Count_{match} refers to the number of overlapping n-grams between the generated and reference summaries.

ROUGE is particularly effective in measuring content fidelity by assessing how much key information is preserved in generated summaries. Its different variants capture various aspects of summary quality: ROUGE-1 tracks unigrams, ROUGE-2 considers bigram matches, and ROUGE-L evaluates the longest common subsequence to account for fluency. However, despite its advantages, ROUGE does not account for semantic meaning, which means two summaries with different wording but identical meaning may still receive a low score. Additionally, since ROUGE prioritizes recall over precision, it tends to favor longer summaries.

3.6.3 BLEU Score

The Bilingual Evaluation Understudy (BLEU) metric [26], originally developed for machine translation, measures how closely a generated summary aligns with reference texts based on n-gram precision.

The BLEU score is computed as:

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3.2}$$

where BP is the brevity penalty, w_n is the weight assigned to each n-gram precision score, and p_n represents the precision of n-grams.

BLEU ensures that generated summaries contain key elements from reference texts while penalizing overly short summaries through its brevity penalty. However, while BLEU effectively captures phrase-level precision, it does not evaluate readability or coherence. As a result, a summary that conveys the correct meaning but uses different wording may receive a low BLEU score.

3.6.4 Readability Evaluation Metrics

Readability metrics help assess whether the generated summaries are understandable to non-expert readers. The following metrics from TextStat were used to evaluate summary readability.

3.6.5 Selected Readability Metrics

- **Flesch Reading Ease:** Measures the readability of text based on sentence length and word syllables. Higher scores indicate easier-to-read text [13].
- **SMOG Index:** Estimates the years of education needed to understand the text. Lower scores indicate more accessible text [21].
- **Coleman-Liau Index:** Computes readability based on the average number of letters and sentences per 100 words. Lower scores suggest greater readability [4].

- **Dale-Chall Readability Score:** Uses a set of familiar words to determine text complexity. Lower scores indicate simpler, more comprehensible language [7].

Table 3.2: Summary of readability metrics and their interpretation.

Metric	**Higher is Better?**	**Interpretation**
Flesch Reading Ease	Yes	Higher = Easier to read.
SMOG Index	No	Lower = Requires less education to understand.
Coleman-Liau Index	No	Lower = More readable for a broad audience.
Dale-Chall Readability Score	No	Lower = Uses simpler and more familiar words.

The evaluation metrics chosen are essential for validating the effectiveness of the summarization models developed in this thesis. They ensure that the summaries are not only faithful to the original texts but also accessible to a wider audience. The next chapter presents the evaluation results using these metrics to compare model performance.

4 Dataset and Model Selection

4.1 Dataset

4.1.1 Description of the Legal Document Dataset

Legal text datasets serve as the foundation for this research, providing the necessary resources to evaluate and improve both summarization and simplification techniques.

For summarization, pretrained models (**T5**, **BART**, and **Pegasus**) fine-tuned on the **BillSum** dataset are utilized. These models, sourced from the LegSum project [6], serve as starting points for further fine-tuning on other datasets to ensure robust performance across diverse legal contexts.

For simplification, the **SUBTLEX-UK corpus** is employed to provide extensive word frequency data, which plays a crucial role in identifying complex words within legal texts [35]. By analyzing word frequency distributions from general English usage, this corpus enables a systematic approach to detecting terms that may pose comprehension challenges for non-expert readers. Additionally, it allows for the development of simplification strategies by comparing the frequency of legal terms against common language benchmarks. This corpus complements the summarization datasets by facilitating a dual focus: enhancing readability through complexity-aware text simplification while maintaining content accuracy in summarization.

These datasets were selected based on the following criteria:

- **Relevance to legal document processing**: They cover legal domains across multiple jurisdictions (US, UK, EU, India).
- **Availability of labeled data**: Ensuring compatibility with supervised learning models.

- **Preprocessed and structured format**: These datasets require minimal cleaning, making them easier to work with and integrate into the NLP pipeline.

4.1.2 Data Sources and Dataset Size

Table 4.1 provides an overview of the datasets used in this research, including their size, composition, and application areas.

Table 4.1: Summary of Dataset Characteristics

Dataset	Source	Documents	Training	Validation	Test
BillSum	US GovInfo (Congressional Bills)	22,218	16,664	2,222	3,332
EurLexSum	EUR-Lex (EU Legal Documents)	1,504	1,128	151	225
GovReport	US Government Accountability Office	19,465	14,598	2,919	1,948
SUBTLEX-UK	English subtitles corpus	37 million tokens	–	–	–
MLS-Long	Civil Rights Cases (US)	4,539	3,404	454	681
MLS-Short	Civil Rights Cases (US)	3,138	2,340	312	486
MLS-Tiny	Civil Rights Cases (US)	1,603	1,207	145	251
InAbs	Indian Supreme Court Judgments	7,150	5,346	713	1,091
UKAbs	UK Supreme Court Judgments	793	595	79	119

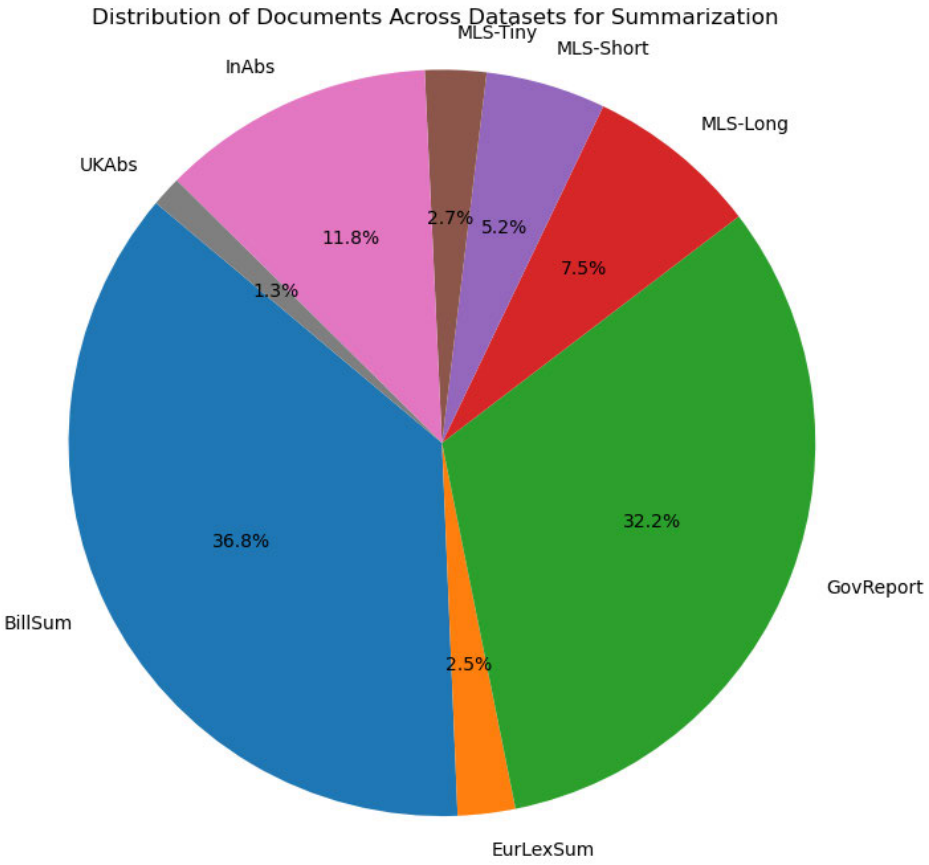

Figure 4.1: Summarization Comparison Analysis

Dataset Length Analysis

To assess the variability in document length, we analyzed the word count and sentence count distributions across datasets. Figure 4.2 and Figure 4.3 present boxplots comparing the distributions.

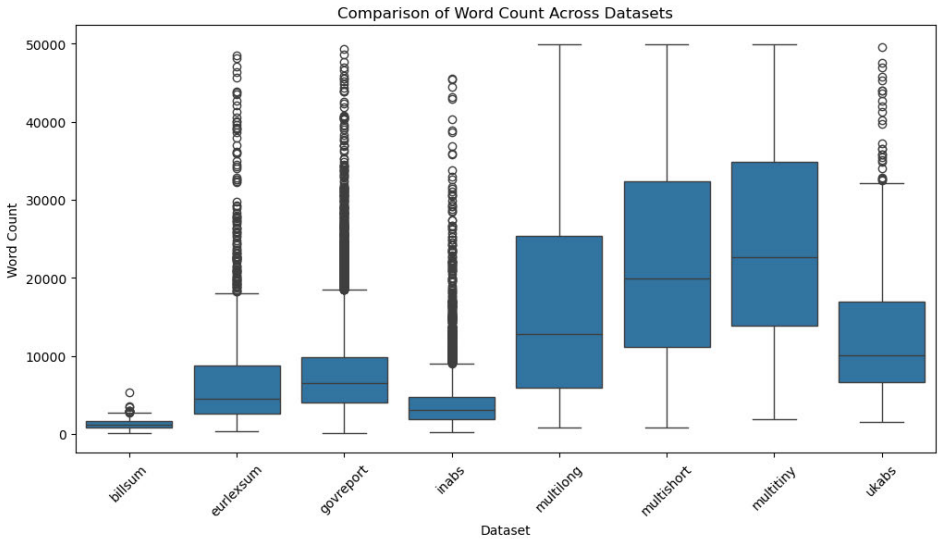

Figure 4.2: Comparison of word count across datasets.

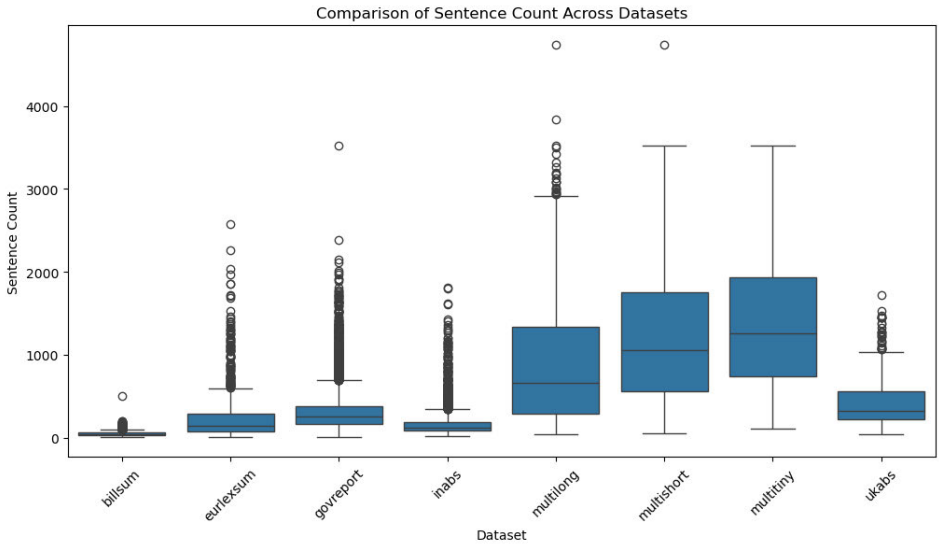

Figure 4.3: Comparison of sentence count across datasets.

The dataset analysis revealed **significant variation** in document lengths:

- The **BillSum dataset** contains the **shortest documents**, with a compact distribution and minimal outliers.

- **EurLexSum, GovReport, and InAbs** exhibit a **moderate spread**, with document lengths ranging between **1,000 and 10,000 words**.
- **MultiLong, MultiShort, and MultiTiny datasets contain the longest legal documents**, often exceeding **30,000 words**, making them the most challenging to process.

Sentence count distribution follows a similar trend, confirming that datasets with **higher word counts tend to have significantly more sentences**. The presence of extreme outliers in **MultiLong and MultiTiny** suggests that some legal texts in these datasets are **considerably longer** than typical documents, possibly entire legal cases or reports.

4.1.3 Simplification Dataset: SUBTLEX-UK Corpus

In addition to the summarization datasets, the **SUBTLEX-UK corpus** was used to support the simplification task [35]. This corpus provides essential frequency data for general English words, enabling the detection of complex terms in legal texts.

Role of SUBTLEX-UK Corpus The SUBTLEX-UK corpus provides valuable word frequency information derived from English subtitles, offering insights into general language usage. In the context of legal text simplification, this corpus helps identify words that are rare in everyday English but frequently appear in legal documents. Additionally, it enables the detection of complex terms based on their low frequency on the Zipf scale, which serves as an indicator of word difficulty. By leveraging this corpus, the simplification process prioritizes replacing complex terms with more accessible alternatives, thereby improving the readability of legal texts while maintaining their intended meaning.

Zipf Score Calculation The complexity of words was determined using Zipf scores, a logarithmic measure of word frequency [40]. Words with low Zipf scores (below a dynamically set threshold) were flagged as complex. Additionally, multi-word expressions specific to legal texts, such as "actus reus" and "prima facie," were included.

Formula: The Zipf Score is calculated as:

$$Z = \log_{10}\left(\frac{f}{N}\right) + 3 \tag{4.1}$$

where:

- f is the frequency of the word in the corpus,
- N is the total number of words in the corpus,
- The constant 3 is added to adjust the scale for interpretability.

Frequency and Context Analysis To enhance detection accuracy:

- **Frequency Comparison**: Words were compared between the SUBTLEX-UK and a legal corpus to identify legal-specific jargon.
- **Supplementary List**: A curated list of complex legal terms ensured comprehensive coverage.

4.1.4 Selection of Documents forthe Experiments

Given computational constraints, this study employs a representative sampling approachrather than using the entire dataset. Instead of processing thousands of documents, I randomly selected 10 documents from each dataset, ensuring a balance between time efficiency, dataset representativeness, and practical feasibility.

The primary reason for this decision is the time required for simplification, which averages 55 minutes per document. With a test set containing **3,105 documents**, fully processing the dataset would take:

$$3,105 \times 55 \text{ minutes} = 170775 \text{ minutes} \approx 118 \text{ days (continuous processing)}.$$

However, many documents exceed the average length, making the actual processing time even longer. Simplifying the entire dataset is therefore impractical. By selecting a random subset of 10 documents, I reduce processing time to a manageable level while still evaluating the model's performance across diverse legal texts.

This approach aligns with research practices in legal NLP. For example, in the **Unsupervised Simplification of Legal Texts** study [2], researchers evaluated simplification models on **500 randomly selected sentences** from a corpus of 27,000 US Supreme Court cases. While their selection focused on individual sentences rather than full documents, the principle remains the same: a well-chosen sample can provide meaningful insights without requiring exhaustive processing. In fact, a single document in this study can contain over 500 sentences, meaning even with just 10 documents, the total number of simplified sentences will significantly exceed 500.

By randomly sampling 10 documents from the entire dataset, this study captures the diversity of legal texts while ensuring the evaluation remains computationally feasible. The documents were randomly selected to prevent any single type of legal document from dominating the evaluation. This method enables an efficient yet broad assessment of the model's ability to process legal texts of varying complexity.

The simplification experiment directly addresses **Research Question 2 (RQ2)** by evaluating whether reducing legal text complexity improves overall readability and summarization effectiveness. By selecting a manageable number of documents for simplification, this study ensures that the experiment remains feasible while still generating insights into the impact of text simplification on legal document processing.

In contrast to simplification, the summarization process is computationally less demanding. Therefore, summarization was performed on the entire dataset, rather than just the selected 10-document subset. This distinction is essential for **Research Question 1**, as evaluating summarization across all available legal documents provides a more comprehensive performance assessment.

In summary, limiting simplification to 10 randomly selected documents ensures that the experiment remains computationally feasible while still providing valuable insights. Meanwhile, evaluating summarization on the full dataset ensures a statistically robust assessment of the summarization model across a diverse range of legal texts.

4.2 Model Selection and Justification

For this study, three candidate models for legal text summarization were evaluated: **Legal-Pegasus**, **BART-BillSum**, and **T5-BillSum**. These models were selected based on their strong performance in previous research on abstractive summarization tasks and

their adaptation to the legal domain. Each model was fine-tuned on the BillSum dataset, making them well-suited for legal document summarization:

- **Legal-Pegasus** (PEGASUS fine-tuned on BillSum) [39]
- **BART-BillSum** (BART fine-tuned on BillSum) [16]
- **T5-BillSum** (T5 fine-tuned on BillSum) [28]

The selection of these models was based on several factors:

- Prior research demonstrating their effectiveness in abstractive summarization tasks [39, 16, 28].
- Their ability to process and generate long-sequence text, which is crucial for legal document summarization.
- Pretraining on the BillSum dataset, making them strong candidates for tasks focused on legal text.

4.2.1 Justification for Model Selection

Previous studies, such as *"Enhancing Legal Document Summarization Through NLP Models: A Comparative Analysis of T5, PEGASUS, and BART Approaches"* [12], have compared T5, PEGASUS, and BART on text summarization tasks. Key findings from this work include:

- **PEGASUS consistently performed best** in abstractive summarization, largely due to its gap-sentence pretraining strategy.
- **BART showed strong extractive summarization capabilities** but was less effective at generating abstractive summaries.
- **T5 demonstrated flexibility** in its approach, but it produced summaries with lower factual consistency compared to PEGASUS.

To ensure these conclusions hold in the legal domain, a validation experiment was conducted (see Chapter 5). In this experiment, all models were fine-tuned on the BillSum dataset and evaluated on legal text summarization tasks. The results confirmed that **Legal-Pegasus outperformed the other models**, reinforcing the findings from prior

research. Based on these results, PEGASUS was selected for further fine-tuning on additional legal datasets to enhance its performance in processing complex legal texts.

5 Experimental Results and Evaluation

This chapter presents the results of the experiments conducted to assess the performance of the models used for legal document simplification and summarization. The evaluation examines multiple aspects of model effectiveness, including:

- **Model selection validation**, verifying whether pretraining on BillSum maintains the performance hierarchy observed in prior research.
- **Summarization-only evaluation**, measuring improvements in PEGASUS performance after additional fine-tuning on domain-specific legal datasets.
- **Simplification-only evaluation**, assessing the extent to which readability is improved through the simplification module.
- **Processing sequence experiments**, investigating whether applying simplification before summarization or vice versa produces better results.

Each experiment is evaluated using a combination of summarization quality and readability metrics. The **summarization quality** is assessed using **ROUGE-1, ROUGE-2, ROUGE-L, and BLEU scores**, which measure how well the generated summaries align with expert references. Meanwhile, **readability and accessibility** are quantified using the **Flesch Reading Ease, SMOG Index, Coleman-Liau Index, and Dale-Chall Readability Score**. These metrics ensure a comprehensive assessment of the effectiveness of both individual components and the full pipeline.

For a detailed explanation of these evaluation metrics and their significance, refer to **Section 3.7 in Chapter 3**.

5.1 Model Selection Validation

To validate that PEGASUS still outperforms BART and T5 in legal summarization, we conducted an evaluation comparing the pretrained models fine-tuned on BillSum. The results are compared against prior findings in [12].

Evaluation Method

The evaluation follows the framework described in Section 5.1, using ROUGE and BLEU scores to measure summarization quality and readability metrics to assess accessibility improvements. Among the ROUGE variants, ROUGE-L is considered suitable for evaluating summarization quality in legal texts, as it captures longest common subsequence (LCS) overlaps, which are crucial for preserving the structure and coherence of complex legal sentences.

Given its importance in assessing how well a model retains essential information, ROUGE-L is emphasized in the analysis, and its distribution is visualized using a box plot in Figure 5.1.

5.1.1 Results

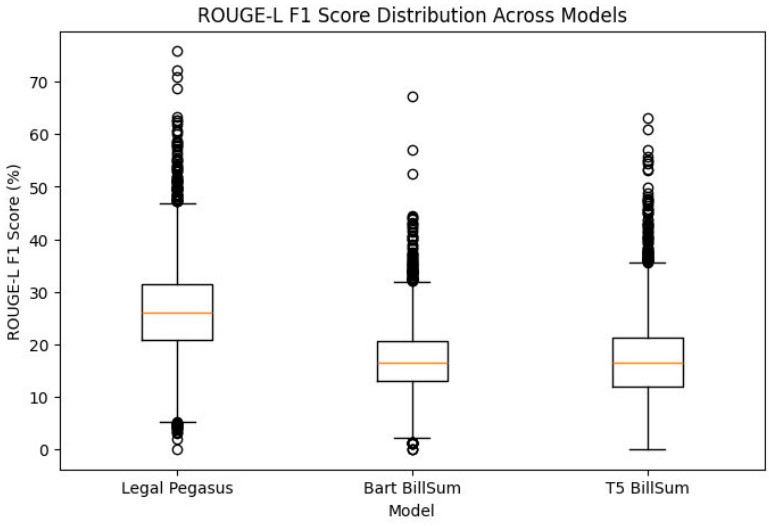

Figure 5.1: ROUGE Score Comparison Across Models

Table 5.1: ROUGE Score Results

Model	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1
Legal Pegasus	28.41	11.09	26.39
Bart BillSum	19.04	5.88	17.17
T5 BillSum	18.14	6.08	16.92

Table 5.2: ROUGE-L F1 Score Summary Across Models

Model	Mean	Std	Min	Max
Legal Pegasus	26.39	9.33	0.0	75.84
Bart BillSum	17.17	6.60	0.0	67.19
T5 BillSum	16.93	8.50	0.0	63.10

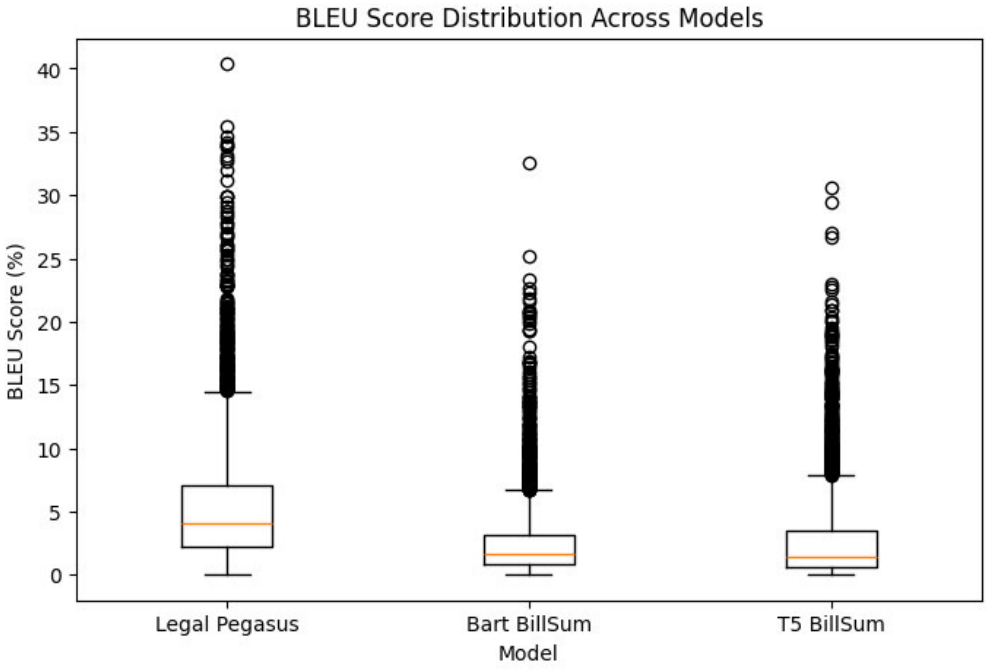

Figure 5.2: BLEU Score Comparison Across Models

Table 5.3: BLEU Score Summary Across Models

Model	Mean	Std	Min	Max
Legal Pegasus	5.53	5.18	0.0	40.38
Bart BillSum	2.56	2.94	0.0	32.54
T5 BillSum	2.76	3.55	0.0	30.61

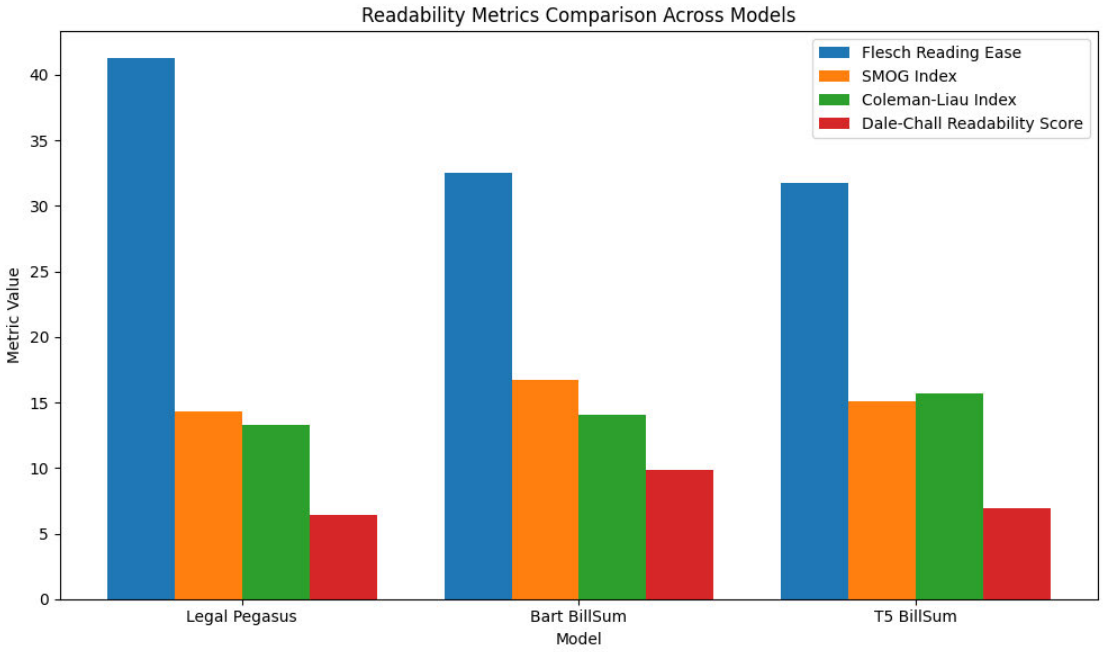

Figure 5.3: Readability Metrics Comparison Across Models

Table 5.4: Readability Metrics Results

Model	Flesch Reading Ease	SMOG Index	Coleman-Liau Index	Dale-Chall Score
Legal Pegasus	41.25	14.32	13.32	6.40
Bart BillSum	32.53	16.68	14.08	9.85
T5 BillSum	31.79	15.04	15.67	6.93

5.2 Summarization Performance: PEGASUS Fine-Tuning

5.2.1 Baseline vs. Fine-Tuned Comparison

The performance of PEGASUS-BillSum is compared against Legal-Pegasus (fine-tuned on legal datasets such as EurLexSum, GovReport, and others).

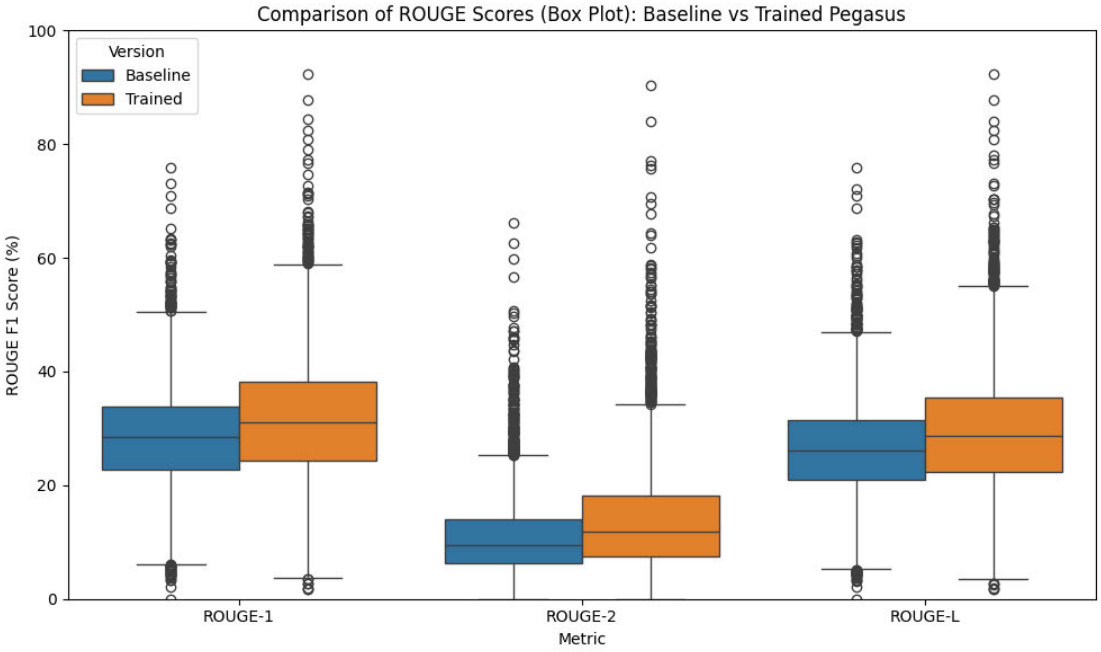

Figure 5.4: Comparison of ROUGE Scores (Box Plot): Baseline vs. Trained PEGASUS

Table 5.5: ROUGE Score Comparison Baseline vs. Fine-Tuned Comparison

Metric	Version	Mean	Std	Max
ROUGE-1	Baseline	28.41	9.70	75.84
ROUGE-1	Trained	31.47	11.86	92.39
ROUGE-2	Baseline	11.10	7.38	66.08
ROUGE-2	Trained	14.26	10.33	90.38
ROUGE-L	Baseline	26.39	9.33	75.84
ROUGE-L	Trained	29.36	11.51	92.39

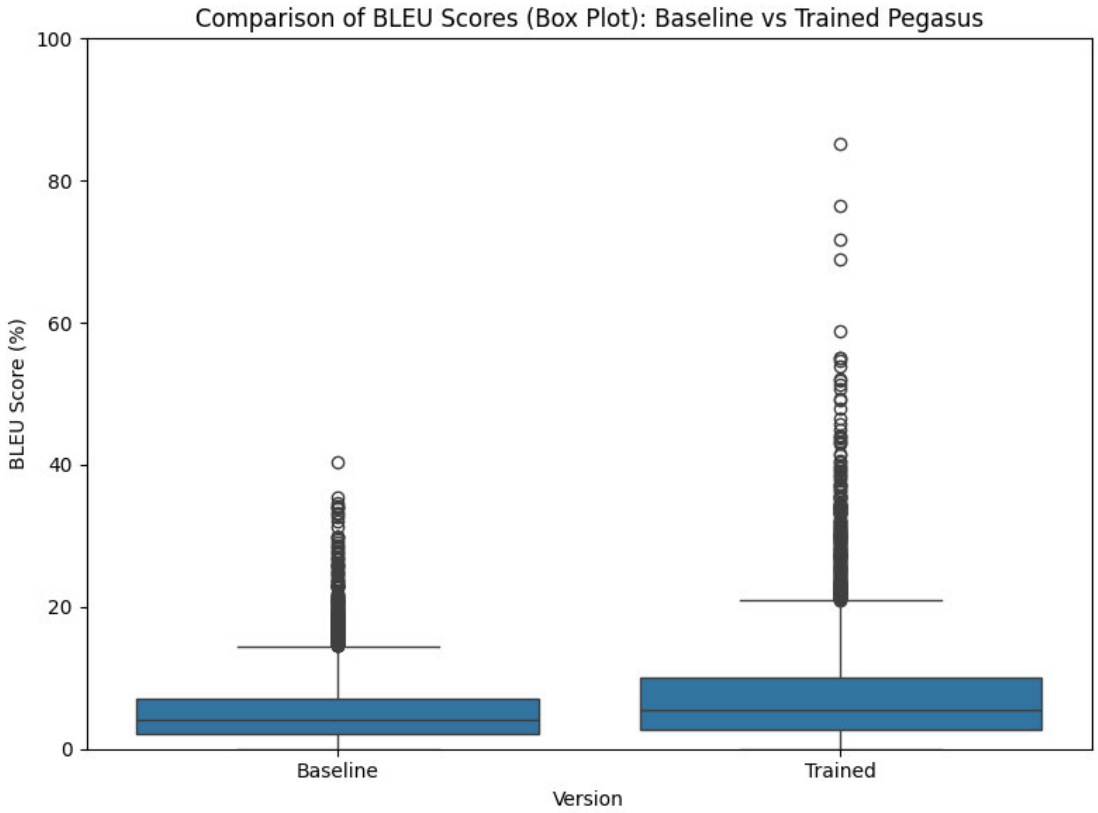

Figure 5.5: Comparison of BLEU Scores (Box Plot): Baseline vs. Trained PEGASUS

Table 5.6: BLEU Score Comparison Baseline vs. Fine-Tuned Comparison

Metric	**Version**	**Mean**	**Std**	**Max**
BLEU	Baseline	5.53	5.18	40.38
BLEU	Trained	7.92	8.25	85.16

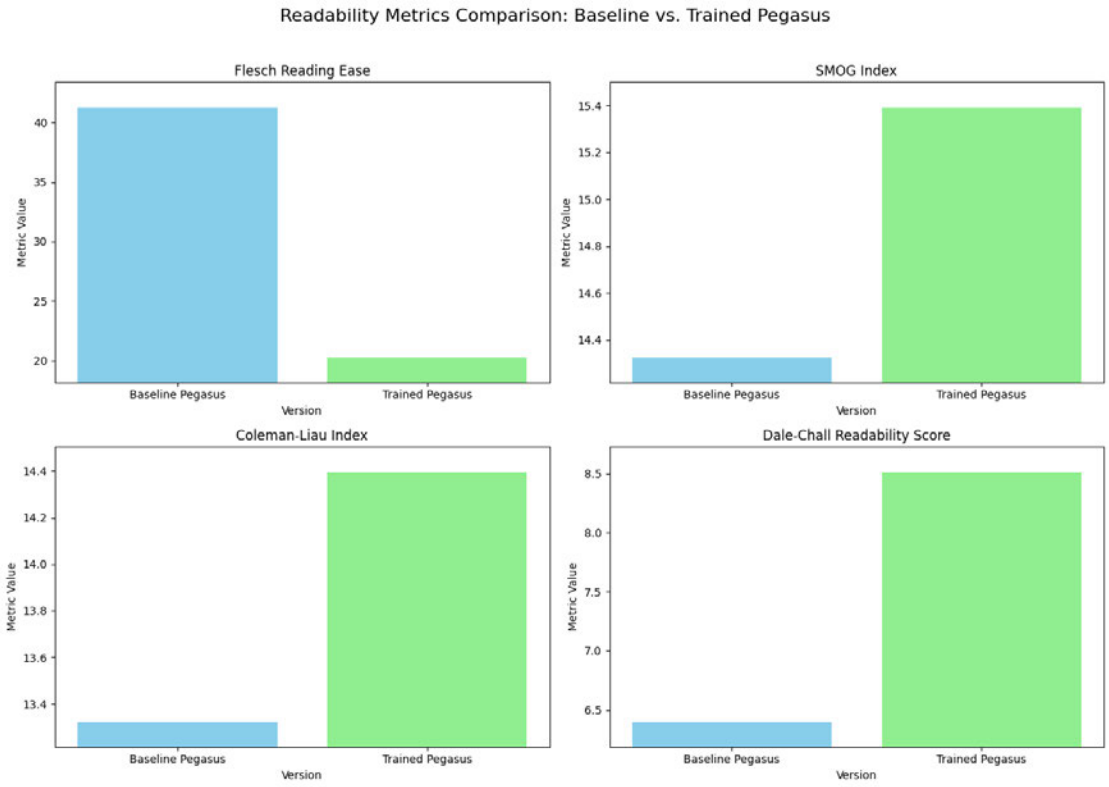

Figure 5.6: Readability Metrics Comparison: Baseline vs. Trained PEGASUS

Table 5.7: Readability Metrics Comparison Across Baseline vs. Fine-Tuned Comparison

Version	Flesch Reading Ease	SMOG Index	Coleman-Liau Index	Dale-Chall Score
Baseline PEGASUS	41.25	14.32	13.32	6.40
Trained PEGASUS	20.28	15.39	14.40	8.51

5.3 Simplification Performance Evaluation

5.3.1 Readability Scores

Comparison of readability metrics before and after simplification.

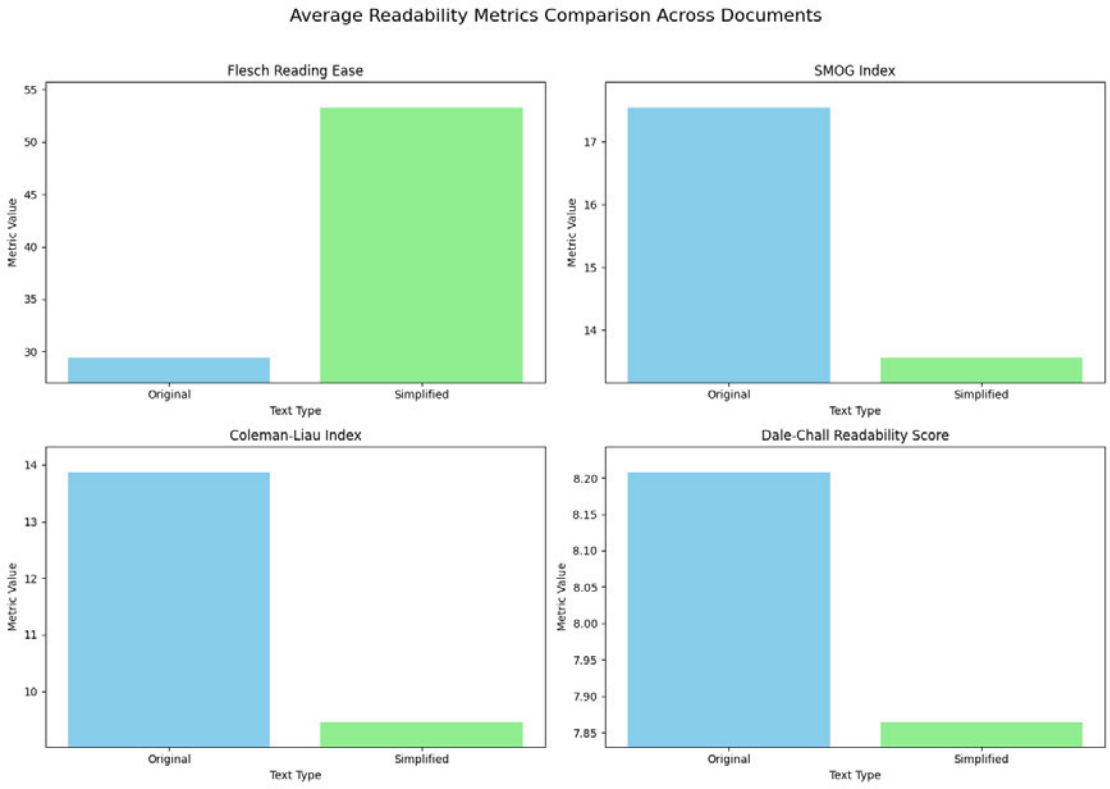

Figure 5.7: Readability Metrics Comparison: Baseline vs. Simplification Process

Table 5.8: Average Readability Metrics Baseline vs Simplification

Version	**Flesch Reading Ease**	**SMOG Index**	**Coleman-Liau Index**	**Dale-Chall Score**
Original	29.45	17.55	13.88	8.21
Simplified	53.30	13.56	9.46	7.87

5.4 Experiment: Evaluating Processing Sequences

This section analyzes the impact of different processing sequences on the final readability and accessibility of legal documents.

5.4.1 Processing Order Comparisons

We evaluate the two possible sequences:

Simplification before Summarization

The legal text is simplified first before being summarized.

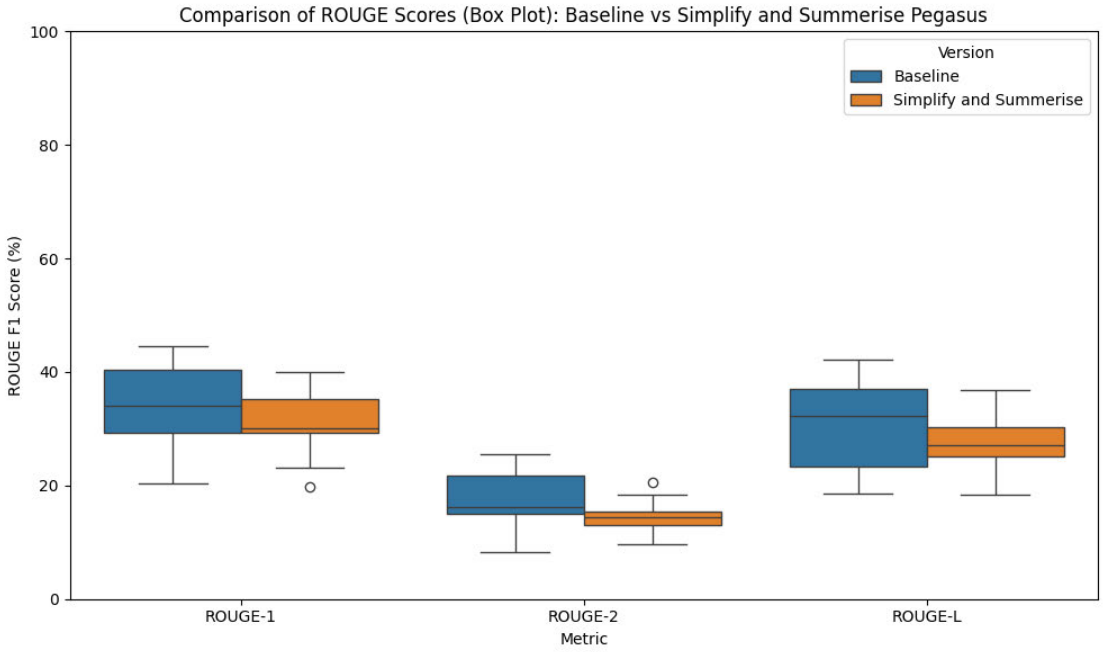

Figure 5.8: Comparison of ROUGE Scores (Box Plot): Baseline vs. Simplify and Summerise PEGASUS

Table 5.9: ROUGE Scores Summary Baseline vs. Simplify and Summerise PEGASUS

Metric	**Version**	**Mean**	**Std**	**Min**	**Max**
ROUGE-1	Baseline	33.78	7.93	20.27	44.53
ROUGE-1	Simplify and Summarise	30.81	6.40	19.85	40.00
ROUGE-2	Baseline	17.27	5.69	8.18	25.49
ROUGE-2	Simplify and Summarise	14.44	3.45	9.71	20.51
ROUGE-L	Baseline	30.54	8.26	18.63	42.11
ROUGE-L	Simplify and Summarise	27.84	6.00	18.32	36.88

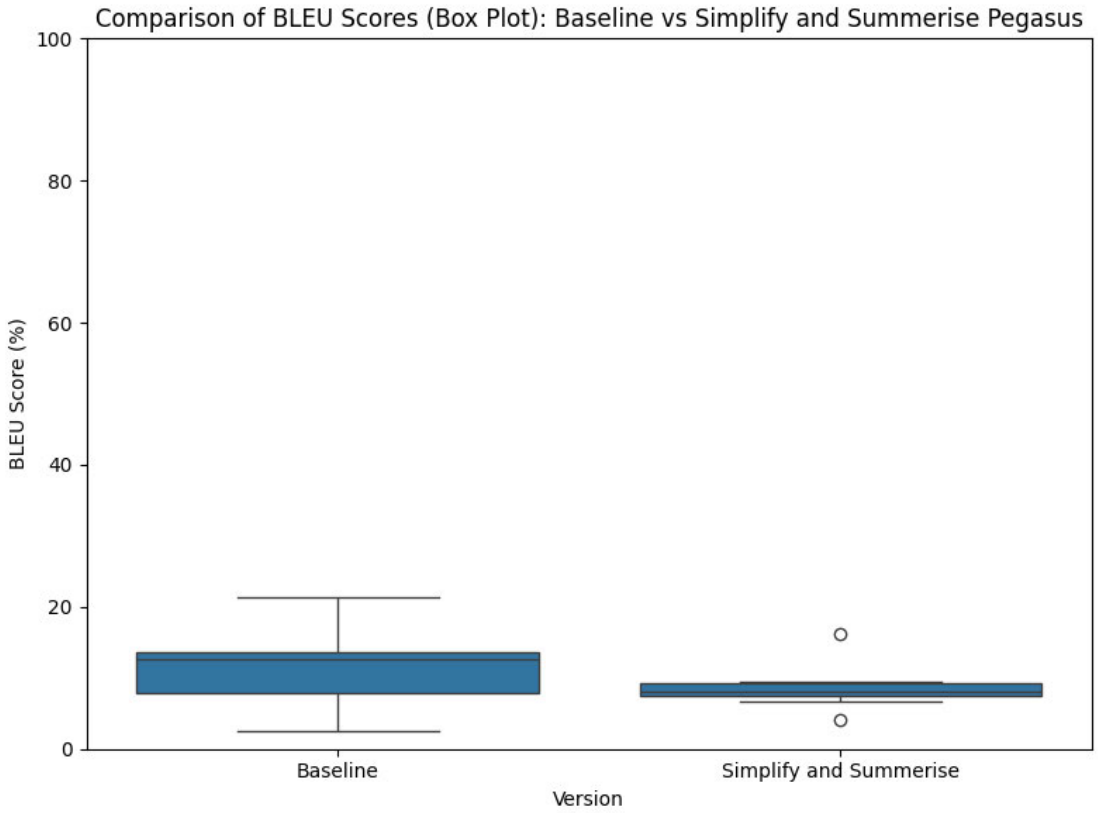

Figure 5.9: Comparison of BLEU Scores (Box Plot): Baseline vs. Simplify and Summerise PEGASUS

Table 5.10: BLEU Scores Summary Baseline vs. Simplify and Summerise PEGASUS

Metric	**Version**	**Mean**	**Std**	**Min**	**Max**
BLEU	Baseline	7.29	6.58	0.17	21.29
BLEU	Simplify and Summarise	7.35	3.61	1.57	16.26

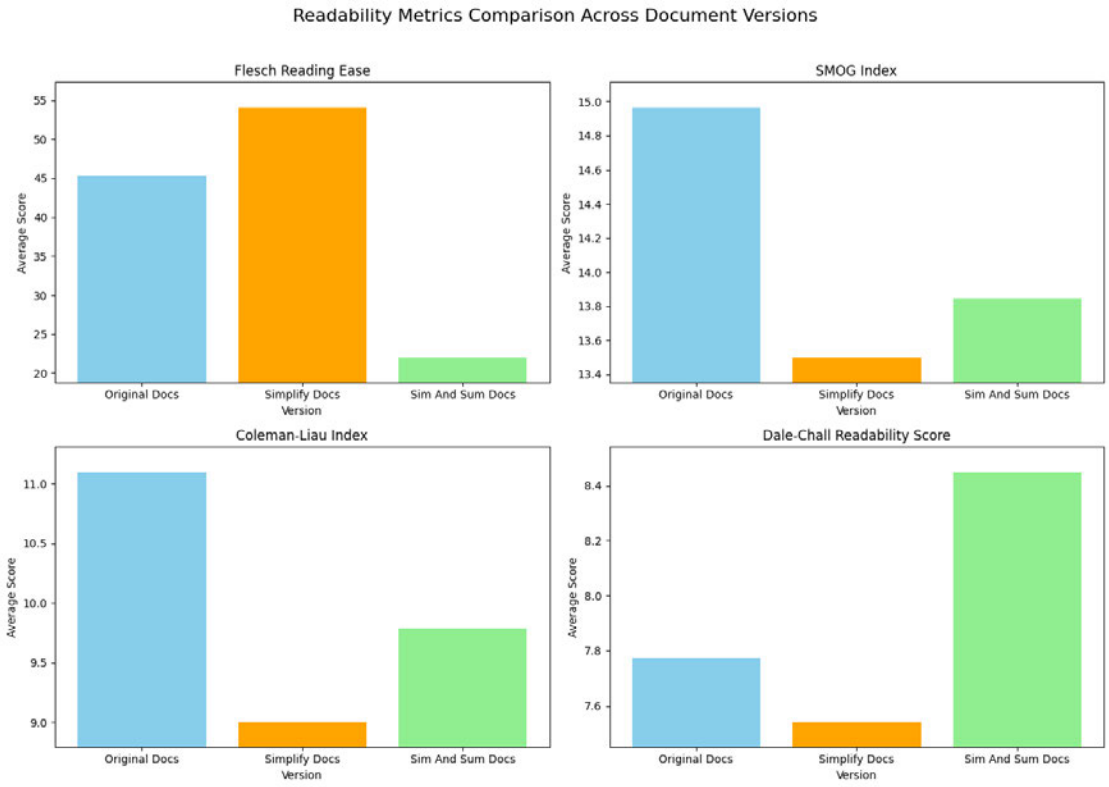

Figure 5.10: Readability Metrics Comparison: Baseline vs. Simplify and Summerise PEGASUS

Table 5.11: Readability Metrics Comparison Across Baseline vs. Simplify and Summerise PEGASUS

Version	Flesch Reading Ease	SMOG Index	Coleman-Liau Index	Dale-Chall Score
Original Docs	45.28	14.97	11.10	7.77
Simplify Docs	54.08	13.50	9.01	7.54
Sim And Sum Docs	22.05	13.84	9.79	8.45

Summarization before Simplification

The text is summarized first and then simplified.

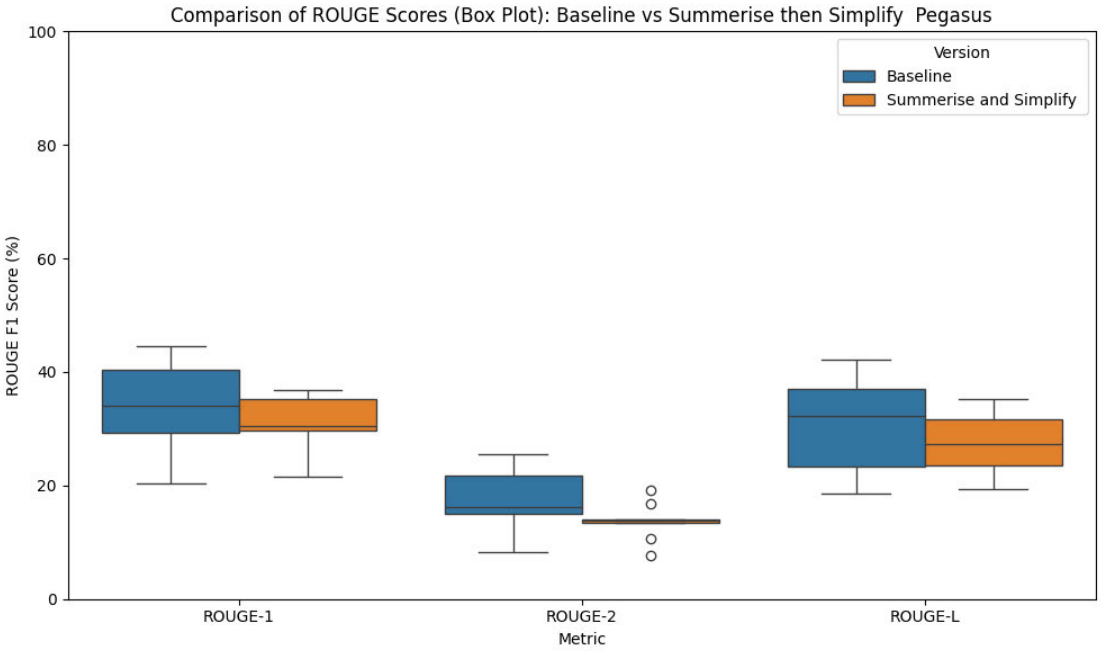

Figure 5.11: Comparison of ROUGE Scores (Box Plot): Baseline vs. Summerise and Simplify PEGASUS

Table 5.12: ROUGE Scores Summary Baseline vs. Summerise and Simplify PEGASUS

Metric	**Version**	**Mean**	**Std**	**Min**	**Max**
ROUGE-1	Baseline	33.78	7.93	20.27	44.53
ROUGE-1	Summerise and Simplify	30.83	5.29	21.55	36.90
ROUGE-2	Baseline	17.27	5.69	8.18	25.49
ROUGE-2	Summerise and Simplify	13.67	3.25	7.64	19.14
ROUGE-L	Baseline	30.54	8.26	18.63	42.11
ROUGE-L	Summerise and Simplify	27.41	5.43	19.34	35.22

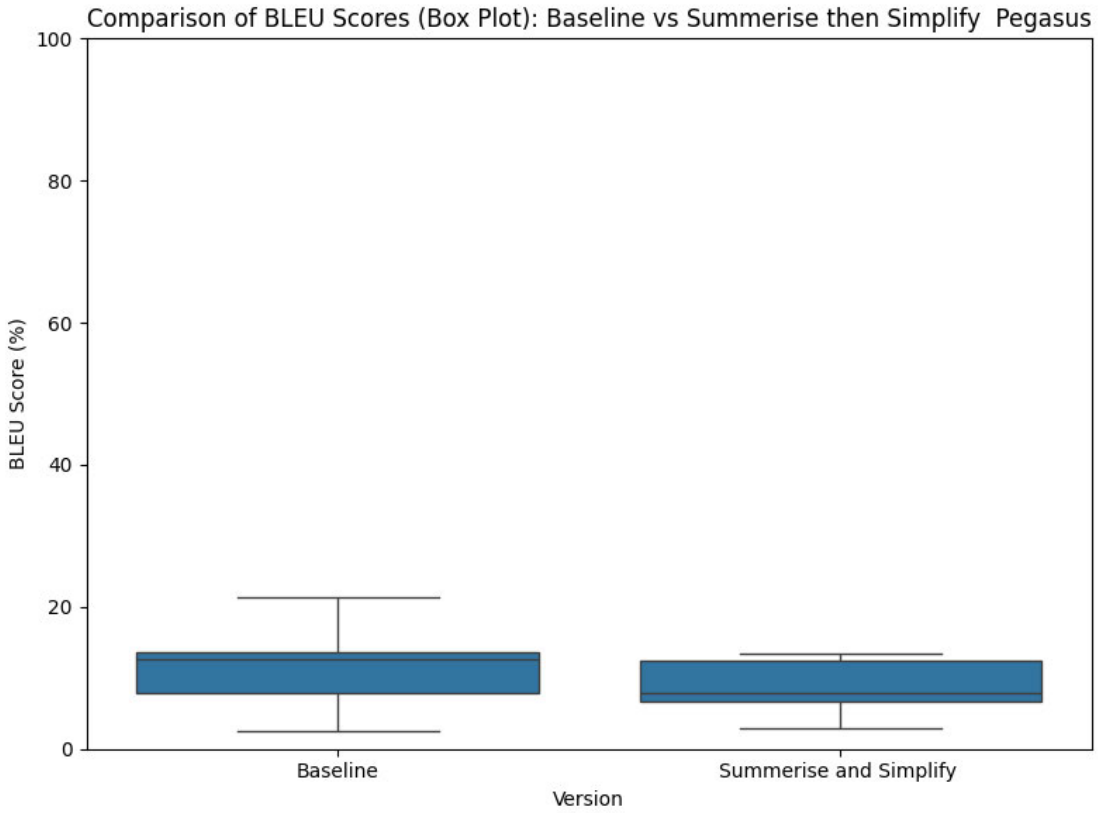

Figure 5.12: Comparison of BLEU Scores (Box Plot): Baseline vs. Summerise and Simplifye PEGASUS

Table 5.13: BLEU Scores Summary Baseline vs. Summerise and Simplify PEGASUS

Metric	**Version**	**Mean**	**Std**	**Min**	**Max**
BLEU	Baseline	5.36	4.58	0.17	13.45
BLEU	Summerise and Simplify	7.35	3.61	1.57	16.26

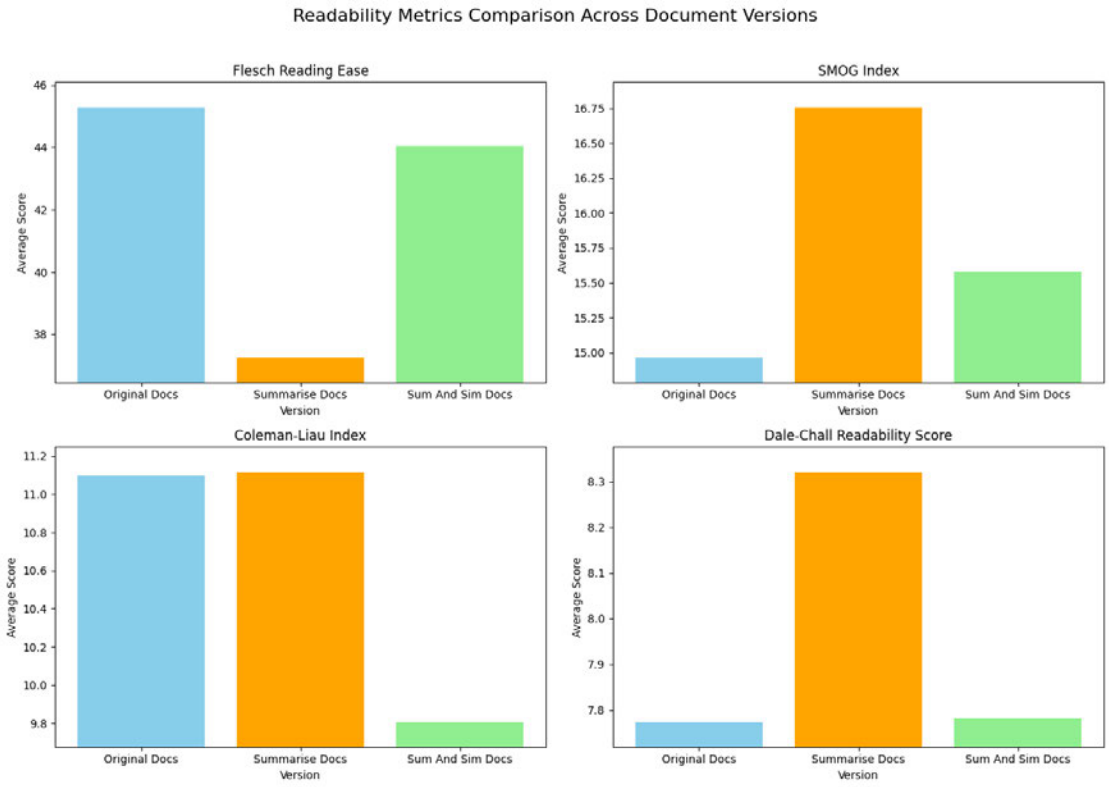

Figure 5.13: Readability Metrics Comparison: Baseline vs. Summerise and Simplify PEGASUS

Table 5.14: Readability Metrics Comparison Baseline vs. Summerise and Simplify PEGASUS

Version	**Flesch Reading Ease**	**SMOG Index**	**Coleman-Liau Index**	**Dale-Chall Score**
Original Docs	45.28	14.97	11.10	7.77
Summarise Docs	37.26	16.76	11.12	8.32
Sum And Sim Docs	44.03	15.58	9.81	7.78

6 Discussion and Analysis

This chapter discusses and interprets the experimental results presented in Chapter 5. The goal is to analyze the effectiveness of fine-tuning pre-trained NLP models for legal document simplification and summarization, evaluate the impact of text simplification on readability, and examine the overall performance of the full pipeline, including different processing sequences.

6.1 Model Selection Validation and Performance Comparison

To validate whether PEGASUS maintains its performance superiority over BART and T5 in legal summarization, we conducted an evaluation comparing the pretrained models fine-tuned on BillSum. The results are compared against prior findings in [12].

6.1.1 Findings from Pretrained Model Evaluation

The experimental results confirm that PEGASUS continues to outperform BART and T5 in summarization quality, as measured by ROUGE and BLEU scores. This finding aligns with prior research, reaffirming PEGASUS's superiority for legal text summarization [12].

A comparative analysis between our dataset and the BillSum-only fine-tuning from prior studies reveals notable differences. Table 6.1 presents a direct comparison of summarization performance before and after incorporating a more diverse legal dataset.

Table 6.1: Comparison of Model Performance (BillSum vs. Our Dataset)

Model	**BillSum Only (Prior Study) [12]**			**Our Dataset (Expanded Legal Corpus)**		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
PEGASUS	**34.25**	**16.63**	**30.22**	**28.41** (-17.1%)	**11.09** (-33.3%)	**26.39** (-12.7%)
BART	26.02	11.87	22.02	19.04 (-26.8%)	5.88 (-50.5%)	17.17 (-22.0%)
T5	32.99	15.52	30.21	18.14 (-45.0%)	6.08 (-60.8%)	16.92 (-44.0%)

The results indicate that, although all models exhibit reduced scores on the expanded dataset, PEGASUS remains the highest-performing model. The performance gap between PEGASUS and the other models persists, validating its effectiveness for legal text summarization across different datasets.

The statistical distribution of scores is illustrated in Figure 5.1 and Figure 5.2, where PEGASUS demonstrates consistently higher median scores and a tighter interquartile range compared to BART and T5.

6.1.2 Impact of Dataset Variability on Model Scores

A key observation from the results is the overall decline in ROUGE and BLEU scores compared to prior research. The primary reason for this performance drop is the increased diversity and complexity of the legal dataset used in this study. Unlike BillSum, which primarily consists of US congressional bills, our dataset includes a broader range of legal documents from different jurisdictions with different ranges of word and sentence counts, leading to higher linguistic variability.

The following insights emerge from this comparison:

- **PEGASUS remains the most robust model** maintaining a significant lead in summarization quality despite the added dataset complexity.
- **BART and T5 experienced more substantial performance drops**, particularly in ROUGE-2 scores, which suggests that their summarization capabilities rely more on dataset specificity.
- **PEGASUS shows lower standard deviation in ROUGE and BLEU scores**, indicating more stable performance across legal documents, whereas BART and T5 exhibit greater variability.

- **Performance extremes highlight dataset complexity**, with minimum scores of zero for all models, but PEGASUS achieving significantly higher peak performance (75.84 ROUGE-L, 40.38 BLEU), demonstrating its ability to summarize well-structured legal content effectively.
- **The observed score reductions are proportional to dataset variability**, reinforcing the hypothesis that summarization models trained on more diverse legal texts face greater challenges in maintaining high performance.

Despite these challenges, PEGASUS consistently delivers the highest summarization scores, confirming its suitability for legal NLP applications. Future research could explore fine-tuning strategies that mitigate dataset-induced performance drops while maintaining the model's robustness across legal domains.

6.2 Effectiveness of Fine-Tuning Pretrained Models for Legal Summarization

The goal of this section is to evaluate the impact of additional fine-tuning on PEGASUS using a broader legal dataset. The evaluation follows the framework described in Section 5.1, using ROUGE and BLEU scores to measure summarization quality and readability metrics to assess accessibility improvements. Since ROUGE-L provides the best alignment with summarization effectiveness, the results of this metric are particularly emphasized in the analysis.

6.2.1 Summarization Performance Improvements

Fine-tuning PEGASUS on an expanded legal dataset led to notable improvements in summarization quality. The following key observations were made:

- **ROUGE and BLEU scores improved across all metrics** after fine-tuning, demonstrating that domain-specific adaptation enhances performance.
- **ROUGE-L scores increased from 26.39% to 29.36%,** confirming an improved ability to capture key legal information.

- **BLEU scores saw a relative increase of 43.2%** (from 5.53% to 7.92%), highlighting improved phrase matching and content fidelity.
- **Higher maximum ROUGE and BLEU scores** indicate that fine-tuning helps PEGASUS generate more accurate summaries for certain types of legal texts.

The statistical distribution of these scores is illustrated in Figure 5.4 and Figure 5.5, where the fine-tuned PEGASUS model demonstrates an overall performance lift.

6.2.2 Comparison of Summarization Metrics

Table 6.2 provides a numerical comparison of summarization performance before and after fine-tuning.

Table 6.2: Summarization Performance Before and After Fine-Tuning (Percentage Change Included)

Metric	**Baseline PEGASUS**			**Fine-Tuned PEGASUS**			**Change (%)**
	Mean	Std	Max	Mean	Std	Max	
ROUGE-1	28.41	9.70	75.84	31.47	11.86	92.39	+10.8%
ROUGE-2	11.10	7.38	66.08	14.26	10.33	90.38	+28.4%
ROUGE-L	26.39	9.33	75.84	29.36	11.51	92.39	+11.3%
BLEU	5.53	5.18	40.38	7.92	8.25	85.16	+43.2%

The results confirm that fine-tuning PEGASUS enhances both content retention (ROUGE scores) and fluency (BLEU scores), making it more effective for legal summarization tasks.

6.2.3 Impact on Readability

In addition to improving summarization quality, fine-tuning also affected the readability of generated summaries. Readability metrics indicate that:

- **Flesch Reading Ease decreased** from 41.25 to 20.28, reflecting a shift towards denser legal language.

- **SMOG Index and Coleman-Liau Index increased**, suggesting a slight rise in sentence complexity.
- **Dale-Chall Readability Score increased** from 6.40 to 8.51, meaning that fine-tuned PEGASUS produced summaries with a higher proportion of uncommon words.

Table 6.3 provides a detailed breakdown of readability metrics before and after fine-tuning.

Table 6.3: Readability Metrics Comparison: Baseline vs. Fine-Tuned PEGASUS (Percentage Change Included)

Version	**Flesch Reading Ease**	**SMOG Index**	**Coleman-Liau Index**	**Dale-Chall Score**
Baseline PEGASUS	41.25	14.32	13.32	6.40
Trained PEGASUS	20.28	15.39	14.40	8.51
Percentage Change	**-50.8%**	**+7.5%**	**+8.1%**	**+32.9%**

While fine-tuning improves summarization accuracy, the increased complexity in readability metrics suggests that additional adjustments (e.g., controlled simplification) may be needed to balance legal accuracy with accessibility.

6.2.4 Qualitative Comparison of Summaries: Baseline vs. Fine-Tuned PEGASUS

This section provides a direct comparison between baseline PEGASUS (pretrained on BillSum) and fine-tuned PEGASUS (trained on the expanded legal dataset). The table below highlights key differences in sentence structure, legal specificity, and content retention.

Table 6.4: Key Sentence Comparisons: Baseline vs. Fine-Tuned PEGASUS

Baseline PEGASUS (Pretrained on BillSum)	**Fine-Tuned PEGASUS (Legal Dataset)**
"The Constitution requires that vacancies in both houses of Congress be filled by special election."	"The Constitution requires that vacancies in both houses be filled by special election; but in the case of the Senate, it empowers state legislatures to provide for temporary appointments until special elections can be scheduled."
"Oregon and Wisconsin are the only states that do not provide for gubernatorial appointments; their Senate vacancies can only be filled by election."	"Oregon and Wisconsin are the only states that do not allow temporary appointments, requiring vacancies to be filled solely by special election."
"A special election was held to fill the vacancy caused by the death of Rep. Patsy Mink."	"If a vacancy occurs between the time of a statewide election and the expiration of the term, the special election winner serves the remainder of the term."
"Some states do not provide for a special election in these cases."	"Some states hold a special election for the balance of the congressional term on the same day as the regular elections."

6.2.5 Impact on Readability and Legal Precision

The qualitative comparison in Table 6.4 highlights how fine-tuning PEGASUS improved the inclusion of legal terminology and procedural details, but at the cost of increased complexity, which aligns with the readability metric results in Table 6.3.

The key takeaways from this comparison are:

- **Legal Formality Increased:** The fine-tuned model adds procedural clarifications (e.g., "state legislatures providing temporary appointments"), improving legal precision but making the text denser.
- **Sentence Length Increased:** While the baseline summary used concise phrasing, the fine-tuned summary expands clauses (e.g., breaking "special election required" into multiple conditions).

- **Complexity Increased, Readability Decreased:** This aligns with the decline in Flesch Reading Ease (41.25 → 20.28) and the increase in Dale-Chall Readability Score (6.40 → 8.51), showing that fine-tuned PEGASUS uses more uncommon legal words.

These findings confirm that while fine-tuning improves legal accuracy and content retention, it also increases syntactic complexity, which can reduce accessibility for non-experts.

6.3 Impact of Text Simplification on Readability and Content Preservation

The results of the text simplification experiment provide clear evidence that the simplification process significantly enhances the readability of legal documents. This section examines the impact of simplification across four key readability metrics: Flesch Reading Ease, SMOG Index, Coleman-Liau Index, and Dale-Chall Readability Score. The numerical results of the experiment are summarized in Table 6.5, while Figure 5.7 provides a visual comparison of readability improvements.

6.3.1 Comparison of Readability Metrics

Table 6.5: Readability Metrics Before and After Simplification

Metric	**Original Text**	**Simplified Text**	**Change (%)**
Flesch Reading Ease	29.45	53.30	+81.0%
SMOG Index	17.55	13.56	-22.8%
Coleman-Liau Index	13.88	9.46	-31.8%
Dale-Chall Readability Score	8.21	7.87	-4.2%

6.3.2 Readability Improvements

After applying the simplification pipeline, **all readability metrics improved**, indicating that the simplified texts are more accessible to non-experts:

- **Flesch Reading Ease Score increased from 29.45 to 53.30 (+81.0%)** This improvement is significant, bringing the texts closer to standard readability levels for general readers. While original legal texts typically score below 30, indicating they are very difficult to read, a score of 53 suggests that the simplified texts are now more accessible and comprehensible.
- **SMOG Index decreased from 17.55 to 13.56 (-22.8%)** A lower SMOG Index indicates that the simplified text requires fewer years of education to understand. The four-point decrease suggests a shift from a complex legal register, typically at the university level, to a more accessible reading level.
- **Coleman-Liau Index decreased from 13.88 to 9.46 (-31.8%)** The Coleman-Liau Index evaluates readability based on sentence length and word complexity. The decrease in the index suggests that the simplification process successfully reduced sentence complexity without significantly altering the core content.
- **Dale-Chall Readability Score decreased from 8.21 to 7.87 (-4.2%)** The Dale-Chall Readability Score uses a list of common words to measure text difficulty. The smaller reduction compared to other metrics indicates that some legal terminology remained unchanged, preserving critical domain-specific content.

  These findings closely align with the results reported in [2], which demonstrated that applying an unsupervised simplification framework significantly enhances readability from legal content. Experimental results in their study confirmed that using domain-specific corpora and transformer-based language models led to greater improvements over general-language simplification methods. The observed increases in Flesch Reading Ease and reductions in SMOG and Coleman-Liau Index further validate that domain-adapted models provide significant readability benefits for legal texts while remaining compatible with downstream legal NLP tasks.

6.3.3 Balancing Readability and Content Preservation

While the results demonstrate a **clear improvement in readability**, it is important to consider whether these gains **came at the cost of content loss**:

  – Legal-specific terms remained largely intact, as shown by the smaller reduction in the Dale-Chall Readability Score.

- The largest changes occurred in structural complexity (sentence length and syntax), rather than in terminology.
- Potential Limitations: Some legal nuances may have been rephrased or simplified, which could impact the legal accuracy of the text.

Overall, the simplification process achieved its primary goal of improving readability. Future refinements could explore controlling simplification intensity to balance ease of understanding with legal precision.

6.4 Impact of Processing Sequence (Simplify First vs. Summarize First)

6.4.1 Impact of Simplification Before Summarization

The experiment applying simplification before summarization demonstrated notable effects on summarization quality and readability. While summarization quality, measured by ROUGE and BLEU scores, showed a slight decline, readability metrics were negatively affected, suggesting that legal complexity was reintroduced during the summarization phase.

Improvements in Summarization Quality

Fine-tuning PEGASUS on simplified texts led to observable differences in summarization performance. Table 6.6 presents the percentage change in ROUGE and BLEU scores.

Table 6.6: Percentage Change in ROUGE and BLEU Scores (Baseline vs. Simplify-Then-Summarize)

Metric	**Baseline**	**Simplify and Summarize**	**Percentage Change**
ROUGE-1	33.78	30.81	**-8.8%**
ROUGE-2	17.27	14.44	**-16.4%**
ROUGE-L	30.54	27.84	**-8.8%**
BLEU	7.29	7.35	**+0.8%**

These results indicate:

- **ROUGE-1 decreased by 8.8%**, indicating a reduction in keyword overlap between summaries and reference texts.
- **ROUGE-2 dropped by 16.4%**, reflecting a decrease in phrase-level abstraction and coherence.
- **ROUGE-L declined by 8.8%**, suggesting a reduction in structural similarity between summaries and references.
- **BLEU remained stable (+0.8%)**, implying that simplification did not negatively affect fluency or n-gram preservation, though it altered content structure.

The decrease in ROUGE scores can be attributed to simplification altering the content's expression, which naturally reduced the n-gram overlap with the reference texts. This does not signify content loss but rather reflects a shift in how information is expressed after simplification, making it less similar in wording to the reference summaries.

Impact on Readability Metrics

Applying summarization after simplification led to significant changes in readability metrics. Table 6.7 presents the percentage change in readability scores.

Table 6.7: Percentage Change in Readability Metrics (Baseline vs. Simplify-Then-Summarize)

Metric	**Simplified Text**	**Simplify-Then-Summarize**	**Percentage Change**
Flesch Reading Ease	54.08	22.05	**-59.2%**
SMOG Index	13.50	13.84	**+2.5%**
Coleman-Liau Index	9.01	9.79	**+8.7%**
Dale-Chall Score	7.54	8.45	**+12.1%**

These results suggest several trade-offs introduced by summarization:

- **Flesch Reading Ease dropped by 59.2%**, indicating a significant increase in the text's complexity and reduced ease of reading.

- **SMOG Index increased by 2.5%**, suggesting that sentence complexity remained relatively high despite simplification.
- **Coleman-Liau Index increased by 8.7%**, highlighting a shift towards more complex sentence structures.
- **Dale-Chall Readability Score increased by 12.1%**, confirming that more difficult or domain-specific words were reintroduced during summarization.

Interpreting the Trade-offs

These results indicate that while simplification improves readability, the summarization step reintroduces legal complexity. This partially reverses readability gains, as evidenced by the drop in Flesch Reading Ease and the increase in the Dale-Chall Readability Score. The summarization model's tendency to prioritize legal accuracy over accessibility results in denser, more formal language.

These findings align with the earlier results in Section 6.2.5, where fine-tuning PEGASUS on legal datasets led to summaries that were more accurate but denser in language. The trade-off between legal precision and readability highlights the need for additional post-processing to maintain accessibility while preserving legal meaning.

6.4.2 Impact of Summarization Before Simplification

The experiment applying summarization before simplification aimed to condense the text first and then simplify it, testing whether summarization could first capture the essential information before applying simplification to improve readability. This sequence of operations yielded mixed results in terms of summarization effectiveness and readability.

Improvements and Changes in Summarization Quality

When summarization was applied first, the PEGASUS model focused on condensing the legal text before simplifying it. Table 6.8 presents the percentage change in ROUGE and BLEU scores.

Table 6.8: Percentage Change in ROUGE and BLEU Scores (Baseline vs. Summarize-Then-Simplify)

Metric	Baseline	Summarize and Simplify	Percentage Change
ROUGE-1	33.78	30.83	**-8.7%**
ROUGE-2	17.27	13.67	**-20.9%**
ROUGE-L	30.54	27.41	**-10.3%**
BLEU	5.36	7.35	**+37.1%**

The following observations can be made from these results:

- **ROUGE-1 decreased by 8.7%**, indicating a slight reduction in keyword overlap between the summaries and reference texts.
- **ROUGE-2 dropped by 20.9%**, which suggests a significant decrease in phrase-level abstraction and coherence.
- **ROUGE-L declined by 10.3%**, reflecting a loss in structural similarity between the summaries and their references.
- **BLEU improved by 37.1%**, suggesting that while content structure changed, fluency and n-gram preservation were enhanced.

Unlike the simplify-then-summarize approach, summarization first led to greater changes in lexical structure before simplification took place. This resulted in a larger drop in ROUGE-2, which measures bigram overlap, indicating that fewer exact phrase matches were preserved from the original text. However, this does not necessarily imply information loss. Instead, the simplification step effectively replaced complex legal terminology with simpler alternatives, naturally reducing direct n-gram matches while maintaining the core meaning of the text.

At the same time, BLEU scores improved significantly, suggesting that the simplification process contributed to grammatical and syntactic consistency, making the final output more fluent. This aligns with our previous findings that ROUGE scores are sensitive to word changes, meaning a lower ROUGE score in this context is a sign that simplification successfully restructured the text rather than omitting critical details.

Impact on Readability Metrics

When summarization was applied before simplification, the readability scores showed a notable improvement in comparison to the baseline, as the summarization process reduced complexity before simplification. Table 6.9 presents the percentage change in readability metrics.

Table 6.9: Percentage Change in Readability Metrics (Summarize-Then-Simplify)

Metric	**Summarized Text**	**Summarize-Then-Simplify**	**Percentage Change**
Flesch Reading Ease	37.26	44.03	+**18.2%**
SMOG Index	16.76	15.58	**-7.0%**
Coleman-Liau Index	11.12	9.81	**-11.8%**
Dale-Chall Score	8.32	7.78	**-6.5%**

These results highlight the following improvements:

- **Flesch Reading Ease increased by 18.2%**, showing that the final text became easier to read after summarization and simplification.
- **SMOG Index decreased by 7.0%**, indicating a reduction in sentence complexity.
- **Coleman-Liau Index dropped by 11.8%**, showing simpler sentence structures in the final output.
- **Dale-Chall Score decreased by 6.5%**, suggesting that fewer difficult words were used.

Unlike the simplify-then-summarize pipeline, this approach successfully improved readability while maintaining content accessibility. The lower SMOG and Coleman-Liau indices indicate that sentence structure became simpler and less dense, aligning with the goal of producing easier-to-read legal summaries.

Interpreting the Trade-offs

These findings indicate that while summarizing first does improve fluency and readability, it leads to a slight reduction in content retention, as reflected in the ROUGE

score decrease. The summarization process seems to prioritize capturing the essential information, but in doing so, it loses some detailed content, which affects the structural and phrase-level quality of the summary.

At the same time, summarizing before simplification results in more readable and fluent summaries. The increase in BLEU score shows that summarization first facilitates grammatical and syntactical improvements, while the simplification step ensures that the text becomes easier to digest.

6.5 Key Insights from Both Processing Orders

The experiments comparing the two different processing orders—simplify-then-summarize and summarize-then-simplify—reveal distinct trade-offs in terms of summarization accuracy, readability, and content retention. By analyzing the ROUGE, BLEU, and readability metrics, we can determine which sequence is more effective for optimizing legal document accessibility and readability.

6.5.1 Comparison of Accessibility Across Processing Orders

Table 6.10 presents the percentage change in summarization performance for both processing orders. The results highlight that simplify-then-summarize yielded slightly higher ROUGE scores, whereas summarize-then-simplify led to a more significant improvement in BLEU scores.

Table 6.10: Comparison of ROUGE and BLEU Score Changes Across Processing Orders

Metric	**Baseline**	**Simplify-Then-Summarize**	**Summarize-Then-Simplify**	**Better Approach**
ROUGE-1	33.78	30.81 (-8.8%)	30.83 (-8.7%)	Similar
ROUGE-2	17.27	14.44 (-16.4%)	13.67 (-20.8%)	Simplify First
ROUGE-L	30.54	27.84 (-8.8%)	27.41 (-10.2%)	Simplify First
BLEU	5.36	7.35 (+37.1%)	7.35 (+37.1%)	Similar

These findings suggest:

– **Simplify-then-summarize** slightly outperformed **summarize-then-simplify** in ROUGE scores, suggesting it leads to better content restructuring by simplifying the text before summarizing.

- **Summarization before simplification** resulted in higher **BLEU scores**, indicating that this approach improves linguistic fluency and coherence in the final output.
- **ROUGE scores** should be interpreted with caution, as they primarily measure word overlap. The lower ROUGE scores in the **simplify-first** approach reflect intentional word replacements due to simplification, rather than a loss in content quality.

6.5.2 Comparison of Readability

Readability scores reveal that the summarization-first approach produced more balanced results, whereas the simplification-first approach led to a denser final text. Table 6.11 highlights these differences.

Table 6.11: Comparison of Readability Scores Across Processing Orders

Metric	Original	Simplify-Then-Summarize	Summarize-Then-Simplify	Better Approach
Flesch Reading Ease	45.28	22.05 (-51.3%)	44.03 (-2.7%)	Summarize First
SMOG Index	14.97	13.84 (-7.5%)	15.58 (+4.1%)	Simplify First
Coleman-Liau Index	11.10	9.79 (-11.8%)	9.81 (-11.6%)	Similar
Dale-Chall Score	7.77	8.45 (+8.8%)	7.78 (+0.1%)	Summarize First

Key observations:

- Summarize-then-simplify retained readability better, as indicated by a much smaller drop in Flesch Reading Ease (-2.7% vs. -51.3%). This suggests that summarization helps maintain accessibility, preventing the text from becoming too dense.
- Simplify-then-summarize resulted in a greater decrease in Flesch Reading Ease, meaning that simplification before summarization led to a more difficult-to-read final text.
- The SMOG Index (which estimates education level required) decreased in the simplify-first approach, indicating that sentence complexity was reduced, making the text more accessible. In contrast, the summarize-first approach led to a slight increase in the SMOG index (+4.1%), meaning the final output was somewhat more complex.

- Both approaches showed similar changes in the Coleman-Liau Index, reflecting that sentence structure complexity was reduced in both cases.
- The Dale-Chall Readability Score, which measures the use of simpler or more familiar words, increased for simplify-then-summarize (+8.8%), suggesting that the simplification process introduced more difficult or domain-specific words, likely due to legal terminology.

6.5.3 Final Evaluation of Processing Order

The choice of processing order—simplify-then-summarize versus summarize-then-simplify—depends on the specific goals of the task:

- Simplify-then-summarize is preferable when the primary goal is summarization accuracy, as it leads to slightly higher ROUGE scores by restructuring content more effectively.
- Summarize-then-simplify is the better option when the focus is on readability, as it avoids excessive complexity, preserving a balance between fluency and accessibility.
- BLEU scores showed minimal difference between the two approaches, suggesting that neither method significantly impacts fluency or syntactic quality.
- The loss in readability in simplify-then-summarize is likely due to the reintroduction of more formal language in the summarization step.

In conclusion, the selection of the processing order should align with the specific objectives: prioritizing summarization accuracy or readability. Both approaches offer unique advantages depending on the context of the legal document processing task.

7 Conclusion and Outlook

7.1 Summary of Objectives and Achievements

This thesis explored how NLP techniques, particularly text simplification and abstractive summarization, can improve the accessibility of legal documents for non-expert audiences. Through extensive experiments, we evaluated pretrained and fine-tuned models, assessed the role of simplification and summarization sequences, and measured their impact on readability and content preservation.

7.1.1 Answering Research Questions

The key findings are summarized as follows:

RQ1: Can NLP-based text simplification and fine-tuned summarization models improve the readability and accessibility of legal documents for non-expert audiences?

The results suggest that while simplification and summarization significantly enhance accessibility, there are notable trade-offs. Fine-tuning models for legal text simplification improves accessibility by making legal content more concise and readable. However, this process also introduces challenges, as fine-tuning tends to reintroduce more formal phrasing and legal terms, which may affect readability. Therefore, while the approach helps preserve legal accuracy, it does not always make the text easier to read for non-expert audiences.

RQ1.1: Do pre-trained summarization models (e.g., PEGASUS, T5, BART) perform optimally for legal text, or does further fine-tuning on additional legal datasets significantly improve their effectiveness in summarization?

The fine-tuning process led to significant improvements in legal text summarization. Among the models evaluated, PEGASUS consistently outperformed BART and T5, proving its superior ability to handle long legal documents. The addition of more diverse legal data during fine-tuning helped PEGASUS maintain its edge, confirming that specialized training improves performance in legal summarization tasks.

RQ2: Can optimizing the sequence and combination of text simplification and abstractive summarization improve the overall readability and accessibility of legal document?

The sequence of simplification and summarization plays a critical role in balancing readability and accuracy. The simplify-then-summarize approach, while maintaining legal terminology, resulted in a slight reduction in summarization quality (ROUGE scores) due to lexical modifications. On the other hand, the summarize-then-simplify approach produced higher readability scores, suggesting that simplification is more effective when applied after summarization.

7.2 Impact on Legal Document Accessibility

This research highlights the potential of NLP models to enhance legal document accessibility by improving summarization effectiveness and reducing linguistic complexity. The choice of processing order has a significant impact on the results, and the best approach depends on the objective:

- For maximizing summarization accuracy and preserving content integrity: **Simplify-then-Summarize** is the preferred approach.
- For optimizing readability and accessibility for non-expert audiences: **Summarize-then-Simplify** proves to be more effective.

7.3 Limitations and Future Directions

Despite its contributions, this study has several limitations:

7.3.1 Limitations

There are several limitations to this study that should be considered. First, due to computational constraints, the models were fine-tuned on a subset of the dataset, rather than the full range of available legal texts. This limitation could have affected the models' generalization ability, and conducting full-scale training with a broader dataset may lead to better performance across a wider range of legal texts.

Another limitation stems from the variability in the legal datasets used in the experiments. While incorporating diverse datasets improved the generalization of the models, it also introduced structural differences across domains. These variations in dataset structure may have impacted the models' ability to adapt uniformly across different legal subdomains, potentially affecting their performance on specific types of legal documents.

Finally, the evaluation metrics used in this study, namely ROUGE and BLEU, emphasize n-gram overlap. While these metrics are widely used for assessing summarization quality, they tend to penalize abstractive summarization methods, which focus on rephrasing content rather than copying it verbatim. This bias in the metrics could have led to an underestimation of the effectiveness of the abstractive models used in this study.

7.3.2 Future Work

There are several avenues for future work that could further improve the results of this study. One potential direction is the addition of a post-processing step to the summarization process. This step would focus on filtering out unnecessary formalities, which could help improve the readability of the summaries and make them more accessible to non-experts.

Another important area for future development is the optimization of the simplification pipeline. Developing more advanced simplification models that strike a

better balance between legal accuracy and readability would provide better results across a wider range of legal documents. This could be achieved by enhancing the models' ability to handle the complexity of legal language while ensuring that the output remains easy to understand.

Furthermore, domain-specific adaptation presents a promising opportunity for future research. By fine-tuning models on specific legal subdomains, such as contract law or criminal law, the models could achieve greater accuracy and become more applicable to specialized legal tasks. This would help tailor the models to the nuances and particularities of different areas of law.

Finally, incorporating human evaluation into future studies will be crucial. Expert evaluations will provide deeper insights into the legal accuracy and usability of the models, complementing the automated metrics used in this study. Human feedback will help refine the models and ensure that they meet the needs of real-world legal applications.

7.4 Implications for NLP and Legal Tech

This research offers significant insights into how NLP can be applied to legal text processing, with a focus on improving accessibility. The results of this study could have important implications for future advancements in legal AI applications:

- **Automated Legal Assistants**: The development of AI-driven legal assistants capable of dynamically simplifying and summarizing legal documents, making them accessible to laypersons.
- **Hybrid NLP Pipelines**: Future systems could integrate rule-based and deep learning approaches to optimize both readability and legal precision.
- **Interactive Legal Document Tools**: Developing tools that allow users to adjust the level of simplification depending on their legal expertise and preference.
- **Cross-Language Application**: The same NLP processes could be transferred to other languages, such as German, to assist individuals who are less familiar with the language. This would provide them with a better chance of

understanding complex legal terms and improve accessibility to legal information.

This study demonstrates both the potential and the challenges of leveraging NLP for legal document accessibility. While simplification and summarization enhance understanding, maintaining a balance between legal precision and readability remains a key challenge for future research. Additionally, applying these techniques to other languages could broaden their impact, helping diverse populations engage with legal content more effectively.

Bibliography

[1] Christopher Adams, Thomas Watson, and John Smith. Longlegalformer: A pretrained long-document transformer for legal nlp. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, December 2022. arXiv preprint. URL: https://arxiv.org/abs/2207.08684.

[2] Mert Cemri, Tolga Çukur, and Aykut Koç. Unsupervised simplification of legal texts, 2022. URL: https://arxiv.org/abs/2209.00557, arXiv:2209.00557.

[3] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, November 2020. Association for Computational Linguistics. URL: https://aclanthology.org/2020.findings-emnlp.261/, doi:10.18653/v1/2020.findings-emnlp.261.

[4] Meri Coleman and T. Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284, 04 1975. URL: https://www.researchgate.net/publication/232574514_A_Computer_Readability_Formula_Designed_for_Machine_Scoring, doi:10.1037/h0076540.

[5] M. Collantes, Maureen Hipe, Juan Lorenzo Sorilla, Laurenz Tolentino, and Briane Paul V. Samson. Simpatico: A text simplification system for senate and house bills. 2015. URL: https://api.semanticscholar.org/CorpusID:115322769.

[6] d0r1h. Legsum, 2023. GitHub repository, accessed: February 2, 2025. URL: https://github.com/d0r1h/LegSum.

[7] Edgar Dale and Jeanne S. Chall. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28, 1948. URL: http://www.jstor.org/stable/1473169.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805, arXiv:1810.04805.

[9] Vladimir Eidelman. Billsum: A corpus for automatic summarization of us legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, page 48–56. Association for Computational Linguistics, 2019. URL: http://dx.doi.org/10.18653/v1/D19-5406, doi:10.18653/v1/d19-5406.

[10] Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nandakishore Kambhatla. Text simplification for legal domain: Insights and challenges. In *Proceedings of the Natural Legal Language Processing Workshop 2022*. Association for Computational Linguistics, December 2022. URL: https://aclanthology.org/2022.nllp-1.28/, doi:10.18653/v1/2022.nllp-1.28.

[11] Ifat Jagirdar, Sakshi Gandage, Bahkti Waghmare, and Iffat Kazi. Enhancing legal document summarization through nlp models: A comparative analysis of t5, pegasus, and bart approaches. 2024. URL: https://ijcrt.org/papers/IJCRT24A3273.pdf.

[12] K. Johnson and D. Patel. Enhancing legal document summarization through nlp models. *International Journal of Creative Research Thoughts (IJCRT)*, 12:810–812, 2024. Accessed: October 20, 2024. URL: https://www.ijcrt.org/viewfull.php?&p_id=IJCRT24A3273.

[13] J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas: (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Chief of Naval Technical Training, Naval Air Station Memphis, Millington, Springfield, 1975. Distributed by NTIS. URL: https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary.

[14] Maria Konnikova et al. The psychology of reading cost cues in online environments. *Journal of Behavioral Decision Making*, 2013. Accessed: October 30, 2024. URL: https://www.readkong.com/page/improving-consumer-comprehension-of-online-contractual-5022580.

[15] Anastassia Kornilova et al. Billsum: A corpus for document summarization of us legislation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2019. URL: https://aclanthology.org/D19-5406.pdf.

[16] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke 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.

[17] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, juli 2004. Association for Computational Linguistics. URL: https://aclanthology.org/W04-1013/.

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL: https://arxiv.org/abs/1711.05101, arXiv:1711.05101.

[19] Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. Chatgpt as a factual inconsistency evaluator for text summarization, 2023. URL: https://arxiv.org/abs/2303.15621, arXiv:2303.15621.

[20] Christopher D Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999. Chapter 11: Probabilistic Context Free Grammars and Chapter 12: Probabilistic Parsing.

[21] G. E. McLaughlin. Smog grading: A new readability formula. *Journal of Reading*, 12(8):639–646, 1969. URL: https://www.jstor.org/stable/40011226.

[22] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In Dekang Lin and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical*

Methods in Natural Language Processing, Barcelona, Spain, juli 2004. Association for Computational Linguistics. URL: https://aclanthology.org/W04-3252/.

[23] George A. Miller. Wordnet: A lexical database for english. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. URL: https://aclanthology.org/H94-1111/.

[24] Islam Nassar, Michelle Ananda-Rajah, and Reza Haffari. Neural versus non-neural text simplification: A case study. In Meladel Mistica, editor, *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, pages 172–177. Association for Computational Linguistics (ACL), 2019. Australasian Language Technology Association Workshop 2019, ALTA 2019; Conference date: 04-11-2019 through 06-11-2019. URL: https://aclanthology.org/U19-1023.pdf.

[25] Joel Niklaus, Lucia Zheng, Arya D. McCarthy, Christopher Hahn, Brian M. Rosen, Peter Henderson, Daniel E. Ho, Garrett Honke, Percy Liang, and Christopher Manning. Flawn-t5: An empirical examination of effective instruction-tuning data mixtures for legal reasoning. 2024. URL: https://arxiv.org/html/2404.02127v1.

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi:10.3115/1073083.1073135.

[27] Lutz Prechelt. *Early Stopping - But When?*, pages 55–69. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998. doi:10.1007/3-540-49430-8_3.

[28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Text-to-text transfer transformer (t5) code and documentation, 2020. URL: https://github.com/google-research/text-to-text-transfer-transformer.

[29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter 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.

[30] Horacio Saggion. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, pages 25–48, 2017.

[31] Omer Shaham, Shira Ronen, and Michael Ben-David. Legalpegasus: Enhancing legal text summarization with domain-specific pretraining. In *Proceedings of the 2023 Conference on Computational Linguistics for Legal Text Processing*. Association for Computational Linguistics, October 2023. arXiv preprint. URL: https://arxiv.org/abs/2303.15621.

[32] Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. Legal case document summarization: Extractive and abstractive methods and their evaluation. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1048–1064, Online only, November 2022. Association for Computational Linguistics. URL: https://aclanthology.org/2022.aacl-main.77/, doi:10.18653/v1/2022.aacl-main.77.

[33] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, pages 1209–1217, 2014. URL: https://www.cs.toronto.edu/~hinton/absps/JMLRdropout.pdf.

[34] The Behavioural Insights Team. Improving consumer comprehension of online contractual terms and privacy policies. Technical report, The Behavioural Insights Team for the Department for Business, Energy and Industrial Strategy, July 2019. URL: https://assets.publishing.service.gov.uk/media/5d3038f0e5274a14f41d289d/improving-consumer-comprehension-online-contractual-terms-technical-report.pdf.

[35] Walter J. B. van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. Subtlex-uk: A new and improved word frequency database for british english. *Quarterly Journal of Experimental Psychology*, 67(6):1176–1190, 2014. PMID: 24417251. doi:10.1080/17470218.2013.850521.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[37] Wikipedia contributors. Boilerplate clause — Wikipedia, the free encyclopedia, 2023. URL: https://en.wikipedia.org/wiki/Boilerplate_clause.

[38] Terry Winograd. Understanding natural language. *Cognitive Psychology*, 3(1):1–191, 1972. URL: https://www.sciencedirect.com/science/article/pii/0010028572900023, doi:10.1016/0010-0285(72)90002-3.

[39] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2020. URL: https://arxiv.org/abs/1912.08777, arXiv:1912.08777.

[40] George Kingsley Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA, 1949.

A Anhang

A.1 Verwendete Hilfsmittel

In der Tabelle A.1 sind die im Rahmen der Bearbeitung des Themas der Bachelorarbeit verwendeten Werkzeuge und Hilfsmittel aufgelistet.

Table A.1: Verwendete Hilfsmittel und Werkzeuge

Tool	Verwendung
LaTeX	Textsatz- und Layout-Werkzeug verwendet zur Erstellung dieses Dokuments
VSCode	Code-Editor für das Schreiben und Debuggen von Skripten
Hugging Face	Nutzung von vortrainierten NLP-Modellen und Datensätzen für das Textverständnis und die Textgenerierung
Haw Hamburg PC Labor	PC-Arbeitsplatz in der Universität für das Ausführen von Experimenten und der Datenverarbeitung
Python	Programmiersprache für die Implementierung und Ausführung der NLP-Modelle
TensorFlow/PyTorch	Frameworks für das Trainieren und Implementieren von Deep-Learning-Modellen
Jupyter Notebooks	Interaktive Umgebung für das Testen und Dokumentieren von Code und Ergebnissen
Git	Versionskontrollsystem zur Verwaltung des Codes und zur Zusammenarbeit
ChatGPT	KI-basierte Unterstützung bei der Verbesserung von Texten, Zitaten und bei der Problemlösung im Rahmen der Forschung

Erklärung zur selbständigen Bearbeitung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

Ort	Datum	Unterschrift im Original