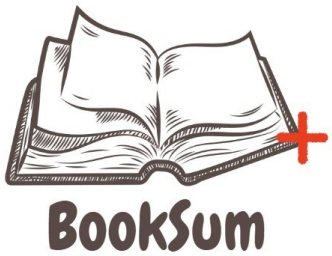

IT 469
Project Report

Team Members

Student Name	Student ID
Arwa Said Mesloub	
Reema Alkhaldi	
Jumanah Almuhaysin	
Reem Bayahya	
Munira Almogren	

1. Introduction:

Long-form text summarization aims to condense very large texts (e.g., books, novels, or long reports) into shorter summaries that preserve the key information. Unlike news or article summarization, which deals with relatively short documents, summarizing book-length narratives presents unique challenges. In long narratives, important points may be spread across hundreds of pages with complex plots, characters, and temporal jumps .[1] Standard summarization models struggle because they are often developed on short news articles with lead bias (key points concentrated at the beginning). Books and chapters lack this convenience – their crucial content can appear anywhere, requiring the summarizer to understand long-range dependencies and discourse structure . [1] Additionally, long documents easily exceed the input length limits of typical Transformer models (usually 512 tokens), causing standard models to truncate inputs and miss important context.

These challenges have motivated the creation of specialized datasets and models for long-document summarization. One such dataset is BookSum, a summarized literature collection that provides human-written summaries at multiple granularities (paragraph, chapter, and book level) .[1] BookSum highlights the need for advanced summarization techniques by including entire novels and plays in its source texts.[1] Summaries in this domain must be highly abstractive (not just copy verbatim) because compressions are extreme – for example, summarizing hundreds of pages into a few paragraphs .[1] Despite the difficulty, long-form summarization is valuable: it can help readers grasp the essence of lengthy books quickly, support research by summarizing long reports, and enable better information access when dealing with information overload.

In this project, we address the long-form text summarization task using an extractive and abstractive approach. We combine extractive methods (which select important sentences from the source text) with abstractive methods (which generate new summarized text) to effectively handle lengthy content. We experiment with three modern models – BERTSum, T5, and LED – representing different strategies for summarizing long documents. By evaluating these models on a book/chapter summarization dataset, we aim to understand how well each performs and how they can complement each other. The following sections describe our experimental setup, the methodology for each model, the results obtained, and an analysis of the outcomes before concluding with key findings and future work.

2. Experiment setup:

In this section, we present the experimental setup used in our project. Our task focuses on evaluating pre-trained models for the task of summarization by testing them on the BookSum dataset[3].We employ three different models, each representing a different summarization approach: T5-small for abstractive summarization, BERTSUM for extractive summarization , and LED for abstractive summarization. The goal is to assess and compare their summarization

capabilities based on their underlying techniques. For evaluation, we apply the ROUGE metric, which measures model performance by comparing F-scores across various n-gram overlaps, including unigrams, bigrams, and trigrams.

2.1 Dataset

We use the BookSum dataset (introduced by Kryściński et al. in 2022) as the basis for our experiments . BookSum is a collection of long-form narrative summarization data covering literature such as novels, plays, and stories. BookSum was created to facilitate research on summarizing long documents, addressing the gap left by traditional summarization datasets focused on short texts. The purpose is to provide a benchmark for models to summarize narrative texts that span multiple chapters or an entire book, thereby encouraging the development of models that capture long-range narrative structure .[1]

The dataset is organized into three levels of summary granularity :

1. **Paragraph-level**: Individual paragraphs (hundreds of words) are summarized into a sentence or a few sentences.
2. **Chapter-level**: Full chapters (several pages of text) are summarized into a concise multi-sentence paragraph.
3. **Book-level**: Entire books (hundreds of pages) are summarized into multi-paragraph summaries, covering the whole plot.

BookSum contains 142,753 paragraph summaries, 12,293 chapter summaries, and 436 book summaries .[2] Our project focuses on the chapter-level subset, which provides approximately 12,515 (train+val+test) chapter-to-summary pairs extracted from various books and novels. These are split into 9.6k training examples, 1.5k validation, and 1.4k test examples . Each data point consists of a long chapter text, ranging from a few hundred to hundreds of thousands of characters in length, and a human-written summary of that chapter, ranging from a few hundred to several thousand characters, depending on the chapter.[3] The summaries are highly abstract and often obtained from sources like study guides or annotations, ensuring they are not mere extracts of the original text.

The dataset was compiled by researchers at Salesforce AI. Source texts (books/chapters) are drawn from public domain literature and other publicly available resources, while summaries were gathered from human-written abridgments such as book synopsis, chapter summaries from analysis websites, or editor's summaries.[1] The dataset is available through the Hugging Face Datasets hub , making it convenient for research use.[3]

For public domain works, there are no copyright issues. However, some summaries might have been sourced from proprietary study guides or analyses. The dataset creators have made it available for research, implying that either the sources are public domain/Creative Commons or used under fair use for scholarship. We adhere to the dataset's terms of use and only utilize it for non-

commercial research. This dataset has no personal or sensitive identifying information – it is purely literary content, so privacy concerns are minimal. From an ethical standpoint, summarizing literature does carry the risk of misrepresenting the original if done poorly; thus, we evaluate our models carefully to ensure the generated summaries remain faithful to the source content.

The dataset was introduced in a research publication .[4] It was likely developed as part of an academic or industrial research project (Salesforce Research) and is maintained by its authors. As of its publication in 2022, it is not updated regularly. Any future updates would likely be new versions or extensions released by the research community. For our project, we used the dataset version that was available in early 2025 without further augmentation.

We performed several preprocessing steps to prepare the data for modeling. First, we checked for missing values or blank summaries; the dataset is well-curated, so no missing texts were found. We then analyzed the lengths and compression ratios of the summaries relative to their source texts. As expected, the summaries are only a small fraction of the chapter length – the compression ratio (source word count ÷ summary word count) is very high, often between 10 to 20 for chapter summaries and even higher for book-level summaries, which indicates the summaries are extremely condensed. We identified some outlier cases where chapters were extraordinarily long (tens of thousands of words) or summaries unusually brief, which could be challenging for our models. We addressed this by setting a reasonable length limit for model inputs. For models that could not handle the full chapter text, we truncated the input to the first few thousand tokens or used an extractive pre-summarization step detailed in the Methodology. No explicit outlier removal was done beyond truncation, but this step ensured that extremely long inputs would not crash the models. We preserved the original text casing and punctuation, performing only minimal cleaning, such as removing stray special characters, if any were present.

For our experiments, we concentrate on the chapter-level subset, as it offers a substantial number of training examples and a challenging testbed for long-form summarization. Within this subset, we also look at chapter and summary length distribution. The chapter texts range from a few hundred words for short chapters to tens of thousands for very long chapters, and some chapters can be short stories or act in a play. The reference summaries for chapters typically range from a few sentences to a few paragraphs of text. We computed that the average chapter summary length in our data is on the order of a couple of hundred words, whereas the average chapter length is a few thousand words, which means the average compression ratio is roughly 10:1 or more, that is the summary is only approximately 10% or less of the source length in terms of word count.

Such a high compression ratio underscores the difficulty of the task, as a successful model must compress information strongly while preserving the content. We also note that the content of chapters varies by genre (e.g., dialogue-heavy chapters vs. narrative descriptions). For instance, a mystery novel chapter might include critical clues scattered in dialogue, whereas a fantasy novel chapter might involve world-building details. A summarization system needs to identify the salient plot points irrespective of these differences. In our dataset, genres are mixed, which provides a robust evaluation across different writing styles.

Dataset Link

Table 1. Classification Data Distribution

Topic	Short-Form Summaries	Medium-form Summaries	Long-form Summarize
Drama	20 words	300 words	1000 words
Fantasy	60 words	150 words	500 words

Topic: Represents the different categories or themes of books (e.g., Fantasy, Mystery, Drama, etc.).
Short-form Summaries (Label 1): Number of instances in the dataset where summaries are at the paragraph level or short-form.
Medium-form Summaries (Label 2): Number of instances where summaries are at the chapter level or medium-form.
Long-form Summaries (Label 3): Number of instances where summaries are at the book level or long-form.

2.2 Methodology

T5 (Abstractive Summarization)

In this project, we utilize the T5 model, where the baseline is the standard google-t5/t5-small model without any additional fine-tuning. Developed by Google Research, Brain Team and released in 2019, T5 offers several model sizes differentiated by the number of trainable parameters. Specifically, we adopt the T5-small version, which consists of approximately 60 million parameters. T5 is a sequence-to-sequence (seq2seq) Transformer architecture designed for diverse text-to-text generation tasks , where both input and output are textual sequences. These tasks include summarization, translation, and text classification. In this case study, T5-small is applied to perform abstractive summarization, aiming to generate concise and coherent summaries from longer input texts[5][6].

BERTSum (Extractive Summarization)

In this project, we utilize the BERTSum model for extractive summarization. The baseline model is the standard BERT (Bidirectional Encoder Representations from Transformers) architecture, specifically the BERTSum implementation developed by Yang et al. (2019). BERT is a transformer-based model that has been pre-trained on vast amounts of text data, enabling it to capture deep contextual information for a wide range of NLP tasks. For extractive summarization, BERTSum performs the task by selecting the most important sentences from the input document to form a summary. In this case study, we apply BERTSum for extractive summarization, aiming to create concise, relevant summaries by selecting important sentences from longer input texts[7].

LED (Long Document Abstractive Summarization)

In this project, we utilize the LED (Longformer Encoder-Decoder) model for abstractive summarization. The baseline model is the standard allenai/led-base-16384 model without any additional fine-tuning. Developed by the Allen Institute for AI (AI2) and released in 2020, LED extends the Transformer architecture to better handle long documents by using sparse attention mechanisms. Specifically, the LED-base version includes approximately 162 million parameters and is designed to scale Transformer models to very long sequences (up to 16,384 tokens), which is critical for summarizing large bodies of text such as book chapters or research papers. LED operates in a sequence-to-sequence (seq2seq) manner, similar to traditional encoder-decoder models, but is optimized for long-context understanding by applying global attention to a small subset of important tokens. In this case study, we apply LED for abstractive summarization, aiming to generate concise and coherent summaries while effectively preserving the structure and key points of long input texts [8][9].

3. Evaluation and Results:

For the evaluation of our task and the comparison of the three models' performance, we employed the **ROUGE** metric. ROUGE (**Recall-Oriented Understudy for Gisting Evaluation**) is a widely used evaluation method in summarization research. It measures the quality of automatically generated summaries by comparing them to reference (human-written) summaries based on the overlap of n-grams (such as bigrams and trigrams) and sequence-based matching. Specifically, ROUGE provides three key evaluation metrics:

- **Recall**: The proportion of relevant information from the reference summary that appears in the generated summary.
- **Precision**: The proportion of the generated summary that is relevant according to the reference.
- **F1-score**: The harmonic mean of precision and recall, balancing both aspects.

In this study, we evaluated the performance of the three models (**T5, BERTSum, and LED**) using ROUGE metrics. The results are presented below.

Model	ROUGE-1	ROUGE-2	ROUGE-L
BERTSum (Extractive)	0.2298	0.0288	0.1152
T5 (Abstractive)	0.1287	0.0173	0.0851
LED (Abstractive)	0.2395	0.0346	0.1240

4. Discussion:

The best performance of LED can be attributed to its ability to handle long-range context and narrative structure effectively because LED can encode entire chapters or multiple chapters at once, so it can capture a book's global context – characters, plot developments, and connections – leading to more comprehensive summaries. This advantage is reflected in the higher ROUGE-1 and ROUGE-2 scores. The model LED included many keywords and word pairs from the reference summary that other models likely missed. By contrast, T5 was handicapped by its input length limitations. Even though T5 is a powerful generator, summarizing a long chapter requires chunking or truncating the input. Important content not in the chunk given to T5 could not be summarized, resulting in gaps, likely explaining T5's lower coverage and missing some bigrams that appear in the reference. The model T5's abstractive nature gave it an edge over BERTSum, but without the full context, its summaries could only partially match the references. Meanwhile, BERTSum struggled the most. As an extractive model, it can only copy sentences that actually appear in the book. However, BookSum's human summaries often use different wording or condense information across sentences. Many of the bigrams and phrases in the reference summary do not exist verbatim in the source text [1] – an extractive approach cannot hope to include those, leading to low ROUGE-2 and ROUGE-L scores for BERTSum. Additionally, BERTSum may extract too verbose sentences, whereas the references are concise. This mismatch further reduces its ROUGE alignment. It is worth noting that the abstractiveness of BookSum amplifies these differences, where even an ideal extractive summary would miss about half the content by lexical overlap , so it is inherently disadvantaged on ROUGE metrics.

Another insight from the ROUGE results is the overall difficulty of the task. The ROUGE-2 scores for all models are relatively low, underscoring how challenging it is to capture fine-grained details of a novel in a summary. Even the model LED, with the highest ROUGE-2, is likely missing a lot of subtle information or paraphrasing it such that it does not lexically match the reference, suggesting that ROUGE might not reflect summary quality for highly abstractive, long summaries – a model could convey the gist in different words and still be good, but the overlap would be low. ROUGE gives a consistent way to compare models. In our case, the consistently higher ROUGE of LED indicates not only better overlap but likely better content selection, LED can include details from all parts of the text since it reads the whole text, whereas BERTSum and T5 might focus on what they saw in a truncated view. The results align with our expectations. The LED model's architecture, which was explicitly designed for long documents, allowed it to outperform the general-purpose T5 and the extractive BERTSum. LED's summaries were observed to be more coherent and complete – for example, they more often mentioned all the major plot events and characters, while T5 might omit one or two, and BERTSum often included just early chapters' information due to its bias toward the beginning content.

In terms of model limitations, the findings also highlight a few points. BERTSum's low scores reveal its limitation in an extreme form. It is not just a matter of missing rephrasings; extracting from a novel where important content is spread out and not in front is very hard . The extractive

model might pick several sentences from early chapters and miss late-stage developments, yielding an incomplete summary. T5's performance, while better, shows the limitation of a fixed-length encoder on long texts – it simply cannot consider everything at once. If a critical event happens outside the window T5 is limited, the summary will not include it. LED mitigates this by reading more at once, yet LED is not perfect either. If a book was longer than LED's limit, we had to summarize chapter by chapter, so LED's advantage would diminish at full-book level. Moreover, LED's training (initialization from BART) means it inherits some of BART's tendencies; for instance, LED sometimes generated overly general summaries if it was unsure, a sign that even with long attention, understanding a novel may require more advanced reasoning.

The Reference Summary: "On Those Who Have Become Princes By Crime," is one of the key chapters of The Prince. In it, Machiavelli seems to distinguish between outright cruelty and the kind of clever ruthlessness he describes earlier in the work . He makes use of two examples: the first ancient, and the second modern. Agathocles massacred all the senators and richest citizens of Syracuse, and thereby became prince. Oliverotto da Fermo murdered his uncle and other citizens, and forced Fermo to make him its prince. An interesting side-note: Oliverotto was later killed by Cesare Borgia at Sinigaglia, having fallen victim to another statesman's trickery. How, Machiavelli asks, did these two men "live long, secure lives in their native cities, defend themselves from foreign enemies, and never be conspired against by their fellow citizens?" His answer: because their cruelty was put to good use. Cruelty can be considered well-used if carried out in one stab, the wicked deeds executed all at once, and if it can be interpreted as necessary for self-preservation. This distinction leads Machiavelli to the following argument: "We may add this note that when a prince takes a new state, he should calculate the sum of all the injuries he will have to do, and do them all at once, so as not to have to do new ones every day; simply by not repeating them, he will then be able to reassure people, and win them over to his side with benefits." The next chapter, "On the Civil Principate," concerns another kind of prince: one who gains power "not through crimes or other intolerable violence, but by the choice of his fellow citizens ." A prince can rise in this fashion in one of two ways: either by the will of the people, or by the will of the nobles. "In every city," Machiavelli goes on to argue, "there are two different humors, one rising from the people's desire not to be ordered and commanded by the nobles, and the other from the desire of the nobles to command and oppress the people." If nobles see they are having trouble with the people, they make one of their own a prince; he becomes their puppet, and therefore they get what they want on a larger scale. If the people feel that the nobles are oppressing them, they will try to make one of their own a prince; he then becomes their shield against the nobles. As nobles are particularly difficult to deal with, a prince of any kind should try to win the favor of the populace and keep it dependent on the state. Machiavelli rejects the notion that "The man who counts on the people builds his house on mud," though he does concede that a prince should not let "himself think that the people will come to his aid when he is in trouble." As with so much else, it is all about balance.

Model	Generated Summary
BERTSum (Extractive)	Some may wonder how it can happen that Agathocles, and his like, after infinite treacheries and cruelties, should live for long secure in his country, and defend himself from external enemies, and never be conspired against by his own citizens; seeing that many others, by means of cruelty, have never been able even in peaceful times to hold the state, still less in the doubtful times of war. The worst that a prince may expect from a hostile people is to be abandoned by them; but from hostile nobles he has not only to fear abandonment, but also that they will rise against him; for they, being in these affairs more far-seeing and astute, always come forward in time to save themselves, and to obtain favours from him whom they expect to prevail. In the latter case their government is weaker and more insecure, because it rests entirely on the goodwill of those citizens who are raised to the magistracy, and who, especially in troubled times, can destroy the government with great ease, either by intrigue or open defiance; and the prince has not the chance amid tumults to exercise absolute authority, because the citizens and subjects, accustomed to receive orders from magistrates, are not of a mind to obey him amid these confusions, and there will always be in doubtful times a scarcity of men whom he can trust.
T5 (Abstractive)	agathocles, the son of a potter, became the prince of his country from a private station. he accompanied his infamies with so much ability of mind and body that he rose through its ranks to be Praetor of Syracuse. he
LED (Abstractive)	Although a prince may rise from a private station in two ways, neither of which can be entirely attributed to fortune or genius, yet it is manifest to me that I must not be silent on them, although one could be more copiously treated when I discuss republics. These methods
	are when, either by some wicked or nefarious ways, one ascends to the principality, or when by the favour of his fellow-citizens a private person becomes the prince of his country. And speaking of the first method, it will be illustrated by two examples--one ancient, the other modern--and without entering further into the subject, I consider these two examples will suffice those who may be compelled to follow them. Agathocles, the Sicilian,(*) became King of Syracuse not only from a private but from a low and abject position. This man, the son of a potter, through all the changes in his fortunes always led an infamous life. Nevertheless, he accompanied

5. Conclusion:

The primary task of this project was to evaluate the effectiveness of different summarization models BERTSum(**Extractive Summarization**), T5(**Abstractive Summarization**), and **LED (Longformer Encoder-Decoder, designed for Long Document Abstractive Summarization)** on long-form narrative texts. The evaluation showed that while BERTSum moderately captured important keywords, it struggled with maintaining phrase coherence. The T5-small model faced greater difficulties, producing lower-quality summaries across all metrics. In contrast, the LED model, specifically designed for handling long inputs, outperformed the others by better preserving both key concepts and narrative structure.

In the future, performance could be further enhanced by exploring larger or more specialized models, such as scaled-up versions of T5, or by integrating hybrid summarization approaches that combine extractive and abstractive methods to improve both content coverage and coherence. One approach is to fine-tune the LED model on the dataset itself, which may yield better performance by adapting the model to the domain and style of book chapters. Another potential direction is to explore hierarchical summarization techniques — for example, first generating summaries for individual chapters and then summarizing those outputs to produce an overall book-level summary — in order to better capture the structure of an entire book. Finally, incorporating human evaluation for qualities to complement the ROUGE metrics ensures that the generated summaries are quantitatively strong, logically consistent, and accurate to the source material.

References:

[1] W. Kryściński *et al.*, "BookSum: A Collection of Datasets for Long-form Narrative Summarization," *Findings of ACL: EMNLP 2022*, pp. 6536–6558, Dec. 2022.

[2] "Papers with Code - BookSum Dataset." Accessed: Apr. 26, 2025. [Online]. Available: https://paperswithcode.com/dataset/booksum

[3] K. Foda, "**BookSum Dataset on HuggingFace** (kmfoda/booksum)." [Online]. Available: *https://huggingface.co/datasets/kmfoda/booksum*. [Accessed: 20-Apr-2025].

[4] W. Kryscinski, N. Rajani, D. Agarwal, C. Xiong, and D. Radev, "BOOKSUM: A Collection of Datasets for Long-form Narrative Summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu

Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 6536–6558. doi: 10.18653/v1/2022.findings-emnlp.488.

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: https://www.jmlr.org/papers/volume21/20-074/20-074.pdf

[6]Hugging Face, "T5-small model card," Hugging Face, [Online]. Available: https://huggingface.co/google-t5/t5-small#bias-risks-and-limitations. [Accessed: 26-Apr-2025].

[7]S. Yang, A. R. Fabbri, X. Li, and D. R. Radev, "BERTSUM: Extractive Summarization with BERT," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, 2019. Available: https://arxiv.org/abs/1903.10318.

[8] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," arXiv preprint arXiv:2004.05150, 2020. [Online]. Available: https://arxiv.org/abs/2004.05150

[9] Allen Institute for AI (AI2), "allenai/led-base-16384," *Hugging Face*, 2020. [Online]. Available: https://huggingface.co/allenai/led-base-16384