WESTERN SYDNEY
UNIVERSITY

Leveraging Large Language Models for Summary/Abstract Extraction of Scientific Texts

Ikhimwin Osakpamwan Emmanuel

22148364

A Report submitted for
INFO 7016 Postgraduate Project A

In partial fulfilment of the requirements for the degree of

Master of Artificial Intelligence.

Principal Supervisor: Dr. Vernon Asuncion

Co-Supervisors: Dr. Manas Patra, Prof. Yan Zhang

School of Computer, Data and Mathematical Sciences

Western Sydney University

October 2025.

ABSTRACT

The exponential growth of scientific publications has intensified the need for automated tools to help researchers efficiently process large volumes of text. Abstracts act as the primary entry point to papers but producing them manually is time-consuming and cognitively demanding. This project explores the use of state-of-the-art large language models (LLMs) for abstract summarisation of scientific literature. Three models were benchmarked: T5-Large, PEGASUS-XSum, and the Longformer Encoder–Decoder (LED).

Performance was assessed using complementary metrics: ROUGE (lexical overlap), BERTScore (semantic similarity), and a novel sentence-level cosine similarity (Top-K) introduced in this study to capture structural alignment. Efficiency indicators, including runtime, memory usage, and token counts, were also analysed. Results show that LED achieved the strongest lexical fidelity but required the most computational resources. PEGASUS was the fastest and most efficient but produced overly concise outputs with weaker coverage. T5 provided the best balance, combining semantic coherence and sentence-level alignment with moderate efficiency.

The findings highlight that no single model dominates across all dimensions, underscoring the importance of multi-metric evaluation. This work contributes a systematic comparison of LLMs for scientific summarisation and provides practical guidance for researchers and institutions deploying automated summarisation systems.

ACKNOWLEDGMENTS

I would like to sincerely thank my supervisors, **Dr. Vernon Asuncion, Dr. Manas Patra, and Prof. Yan Zhang**, for their guidance, encouragement, and invaluable feedback throughout the course of this project. Their expertise and support have been instrumental in shaping the direction and quality of my work.

I am also grateful to the **School of Computer, Data and Mathematical Sciences at Western Sydney University** for providing the resources and academic environment necessary to complete this research.

Lastly, I wish to thank my family and friends for their patience, understanding, and unwavering support during my studies.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER I: INTRODUCTION

The rapid growth of publications in computer science has reached a point where even active researchers struggle to keep up with the pace (Bornmann et al., 2021). Abstracts have become the first and often decisive point of contact between a reader and a paper, guiding whether the full text is worth exploring. This project is motivated by the need to automate the creation of meaningful, concise, and representative summaries of scientific papers, reducing the cognitive and time burden on researchers.

Summarization is the process of condensing a long piece of writing into a shorter version that preserves the core ideas and information. Broadly, there are two approaches: extractive summarization, which selects key sentences or phrases directly from the source text, and abstractive summarization, which generates new sentences that paraphrase the original content. Extractive methods closely follow the wording of the source but often lack cohesion, while abstractive methods demand a deeper understanding of content and language generation (Nenkova and McKeown, 2011). With the advent of Large Language Models (LLMs), abstractive summarization has become increasingly feasible, particularly for specialized domains such as scientific research (Zhang et al., 2024).

This project focuses on leveraging state-of-the-art LLMs to generate high-quality summaries of scientific documents. Three leading models were selected to represent complementary strengths in abstractive summarization: Text-to-Text Transfer Transformer (T5)-Large (Raffel et al., 2019), which has demonstrated strong general performance across text-to-text tasks; Pre-training with Extracted Gap-sentences for Abstractive SUmmarization (PEGASUS)-Xsum (Zhang et al., 2019), a model designed

with a pretraining objective tailored to summarization; and LED (Longformer Encoder-Decoder) (Beltagy et al., 2020), which is optimized for handling long input sequences typical of scientific writing.

Summarising specialised documents such as scientific, financial, and legal texts is more difficult than summarising general content because of their complexity and domain-specific language. Scientific papers often contain technical terminology, mathematical expressions, and structured rhetorical patterns that must be preserved when condensed (Cohan et al., 2018). Financial reports and legal documents rely on precise, formal language, where even small omissions or paraphrasing errors can significantly alter meaning or interpretation (Chalkidis et al., 2021). Generic summarisation approaches that work reasonably well on news or conversational text often fail in these contexts, producing outputs that may be fluent but incomplete or inaccurate.

The stakes are also higher in specialised domains. An inaccurate scientific summary may omit important results; errors in financial summarisation can misinform investors and regulators; and mistakes in legal summarisation may lead to serious misinterpretations (Zhong et al., 2020). These challenges underline the limitations of purely extractive techniques and highlight the need for advanced abstractive methods that balance conciseness with fidelity. This motivates the focus of this project on leveraging large language models for scientific text summarisation, where long input sequences and semantic accuracy are critical.

A big challenge in summarising specialised texts is the limitation on token capacity in current large language models. Scientific, financial, and legal documents are often very

long, frequently exceeding thousands of words. Many transformer-based models, however, have strict maximum input lengths (e.g., 512 or 1024 tokens for standard models), which means that only a fraction of the text can be processed at once (Beltagy et al., 2020). This limitation increases the risk of information loss, since critical parts of the document may be truncated or ignored. While models like the Longformer Encoder-Decoder (LED) have introduced sparse attention mechanisms to extend the input window, long documents in specialised domains still pose difficulties in capturing both global context and fine-grained details.

It is also worth noting that a number of specialised AI tools already exist for summarisation and research assistance, many of which are designed for commercial use. Tools such as SciSummary, Scholarcy, and AI features integrated into platforms like Elicit or Semantic Scholar offer researchers automated ways of condensing papers or extracting highlights. Similarly, financial and legal domains increasingly use proprietary summarisation systems embedded within enterprise solutions for regulatory compliance, contract analysis, or market intelligence. However, these tools are often closed-source, subscription-based, and domain-restricted, limiting transparency and generalisability. This underscores the value of academic studies like this project, which seek to evaluate summarisation models systematically in the scientific domain and contribute to open knowledge on their strengths, limitations, and applicability.

The study begins by constructing a dataset of 1,000 scientific papers sourced from arXiv in the categories of Artificial Intelligence (cs.AI) and Machine Learning (cs.LG). Each paper provides an introduction section as input and its abstract as the gold reference for

evaluation. A fixed subset of 15 papers is used to benchmark the three models under controlled conditions. Model performance is assessed with Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004), which measures lexical overlap between generated summaries and abstracts, and Bidirectional Encoder Representations from Transformers (BERTScore) (Zhang et al., 2019), which evaluates semantic similarity using contextual embeddings. Efficiency and memory usage are also examined to highlight trade-offs between quality and computational cost.

The broader aim of this work is to identify which model is most effective for scientific summarization and to fine-tune the strongest candidate on a larger dataset of 1,000 papers. In doing so, the project provides insights into both the strengths and limitations of applying general-purpose summarization models to the scientific domain. Ultimately, the outcomes are expected to bridge the gap between generic summarization systems and the specific needs of researchers, offering a more reliable and efficient way to process and digest the growing body of scientific literature.

Research Objectives

The objectives of this project are:

- **Benchmarking**: To benchmark these models; **T5-Large**, **PEGASUS-XSum**, and **LED** on the task of scientific introduction-to-abstract summarization.
- **Evaluation**: To evaluate model performance using lexical, semantic, and sentence-level similarity metrics.

- **Efficiency**: To compare efficiency across models in terms of runtime and memory usage.
- **Identification**: To identify the most effective model for fine-tuning on a large-scale dataset of 1,000 papers.

Report Structure

The remainder of this report is organised as follows: Chapter II reviews related literature, Chapter III describes the methodology, Chapter IV presents the results, Chapter V discusses the findings, and Chapter VI concludes with a summary and outlines the research plan for Project B.

CHAPTER II: Literature Review

The task of text summarisation has long been a central problem in Natural Language Processing (NLP), aiming to condense long-form text into a concise version while preserving essential meaning. Early approaches were primarily extractive, identifying and concatenating the most salient sentences from the source document. Although extractive methods are computationally efficient and maintain fidelity to the original text, their outputs often lack coherence and fail to capture deeper semantic relationships. Abstractive summarisation, by contrast, seeks to paraphrase the source text and generate new sentences that convey the key ideas (Nenkova & McKeown, 2011; Rush et al., 2015; See et al., 2017). This approach more closely resembles human summarisation but demands advanced language generation capabilities.

The introduction of transformer-based architectures revolutionised abstractive summarisation. The self-attention mechanism of the Transformers (Vaswani et al., 2017) Transformers greatly simplified sequence processing in NLP and allowed models to capture long-range dependencies across text, providing the foundation for modern LLMs. Within this paradigm, several models have been particularly influential. The T5 reframed all NLP problems as text-to-text tasks, enabling a unified architecture that could be fine-tuned for summarisation (Raffel et al., 2019). PEGASUS advanced the field with a pretraining objective specifically designed for summarisation: gap-sentence generation, which encourages the model to predict removed sentences from the surrounding context (Zhang et al., 2019). This objective aligns closely with the summarisation task and has led to state-of-the-art performance on multiple benchmarks. Meanwhile, the LED

extended transformers to handle long documents by incorporating sparse attention patterns, allowing inputs far beyond the 512-token limit of standard models (Beltagy et al., 2020). This makes LED particularly suitable for summarising scientific texts, where introductions and related sections are often lengthy.

Evaluation remains a critical dimension of summarisation research, as it determines whether system-generated summaries are genuinely useful to humans. For many years, ROUGE (Lin, 2004) has been the dominant metric, measuring the overlap of n-grams and longest common subsequences between system outputs and reference abstracts. While widely adopted for its simplicity, ROUGE focuses only on surface-level word matches, which means it can undervalue summaries that use valid paraphrases or alternative wording. To overcome these limitations, BERTScore was proposed as a semantic evaluation metric that leverages contextual embeddings from pretrained language models to compute similarity between candidate and reference sentences (Zhang et al., 2020). This allows recognition of paraphrases and semantic equivalence, providing a more nuanced measure of quality. Recent studies have increasingly combined ROUGE with BERTScore to balance lexical precision with semantic fidelity (Fabbri et al., 2021).

More recently, the emergence of advanced LLMs such as GPT-4 has highlighted both the strengths and shortcomings of traditional metrics, since these models can generate highly fluent outputs that may not be fully captured by lexical or embedding-based scores alone. This has motivated ongoing research into LLM-as-a-judge approaches, where large models are themselves used to provide human-aligned evaluations of summary quality (Liu et al., 2023).

Summarisation of scientific text poses distinct challenges compared to news or conversational data. Scientific papers are longer, more technical, and follow structured rhetorical conventions, typically including background, problem definition, contributions, and results (Cohan et al., 2018). While large summarisation datasets exist for domains such as news (CNN/DailyMail, XSum), scientific datasets are less abundant, and the complexity of technical language exacerbates the limitations of standard models. Cohan et al. (2018) introduced the PubMed and arXiv datasets for long-document summarisation, highlighting that traditional encoder-decoder architectures struggle to capture the breadth of information required for effective scientific abstracts. Models like LED have demonstrated improved performance on these datasets, but systematic comparisons across multiple LLMs for the task of summarising scientific introductions remain limited.

This project builds upon the strengths and limitations identified in prior work (Cohan et al., 2018; Beltagy et al., 2020). By benchmarking T5-Large (Raffel et al., 2019), PEGASUS-XSum (Zhang et al., 2019), and LED (Beltagy et al., 2020) on scientific introductions from the arXiv cs.AI and cs.LG categories, it provides a controlled evaluation of three leading abstractive models. Unlike prior research, which has focused either on general-domain summarisation or on biomedical texts (Cohan et al., 2018), this work directly investigates the application of LLMs to artificial intelligence and machine learning literature. The combination of ROUGE and BERTScore ensures that both lexical and semantic aspects of summary quality are captured, while measurements of runtime and memory usage introduce a practical perspective that is often overlooked in the literature. Ultimately, this review positions the project at the intersection of abstractive

LLM-based summarisation, long-document modelling, and domain-specific evaluation, aiming to close an important gap in how researchers interact with the rapidly expanding body of scientific publications.

However, both ROUGE and BERTScore evaluate summaries primarily at the lexical and token levels. While ROUGE is limited to surface-level n-gram overlaps and BERTScore leverages contextual embeddings to capture semantic equivalence, neither metric explicitly accounts for structural alignment at the sentence level. In the context of scientific abstracts, where key contributions and claims are often distributed across distinct sentences, evaluating only at the word or token level may obscure whether the generated summaries preserve the most salient units of meaning. To address this limitation, this study introduces an additional evaluation metric based on sentence-level cosine similarity. By embedding individual sentences from both the reference abstract and the generated summary and then identifying the strongest semantic correspondences, this metric provides a complementary perspective on how effectively each model captures and condenses the core propositional content of scientific texts. Incorporating this measure alongside ROUGE and BERTScore ensures a more holistic evaluation framework, balancing lexical precision, semantic fidelity, and structural coverage.

CHAPTER III: METHODOLOGY

3.1 Dataset and Sampling Source

Papers in cs.AI (computer science Artificial Intelligence papers) and cs.LG (computer science Machine Learning papers) were obtained from arXiv. For each paper, the Introduction serves as the model input and the authors Abstract serves as the reference (target) summary (Cohan et al., 2018). The full corpus contains 1,000 papers. A fixed subset of 15 papers is used for controlling benchmarking of the models, so the comparisons are the same for the 3 models.

The 15-paper subset was sampled from the 1,000-paper pool using a fixed random seed (42) to ensure reproducibility.The same arxiv_id list is reused for every model.

Data format. Each model's outputs are stored as JSONL (one record per paper) with the following fields:

- arxiv_id, title
- reference_abstract
- generated_summary
- model_name
- time_sec, gpu_mem_bytes
- input_tokens, output_tokens

Data Preparation for Evaluation

The model outputs were consolidated into three DataFrames, one for each model (LED, PEGASUS, and T5) (Beltagy et al., 2020; Zhang et al., 2019; Raffel et al., 2019). Each DataFrame contained the paper metadata (arXiv ID, title), the gold reference abstract, the

generated summary, and efficiency statistics (runtime, memory usage, input/output tokens). To ensure fair comparability, the DataFrames were filtered to the same 15 consistent papers, which formed the basis for all downstream analyses presented in Chapter IV.

Rationale for Focusing on 15 Papers

Although all three models were initially run on 25 arXiv papers, not all generated summaries were of sufficient quality for a balanced comparison. Some outputs were unusually short, others appeared truncated when the introduction text exceeded the model's input window, and in a few cases LaTeX-heavy introductions inflated token counts or disrupted sentence segmentation. To prevent these anomalies from biasing the evaluation, the analysis was restricted to the 15 papers where all models produced reasonably consistent and representative summaries. This filtering ensured that the reported results reflect genuine differences in summarisation performance rather than artefacts of tokenisation limits or noisy inputs.

The complete 15-paper benchmark tables for each model are provided in **Appendix A (Tables A1–A3)**. These include the reference abstracts, generated summaries, and efficiency statistics, and serve as the raw outputs underlying the ROUGE, BERTScore, Top-K similarity, and efficiency evaluations reported in Chapter IV.

3.2 Pre-processing

Prior to model inference, a minimal pre-processing pipeline was applied to ensure that the input texts and reference summaries were clean and consistent across all papers. Each paper's Introduction and Abstract were first extracted and saved in a structured format.

Non-content artefacts such as excess whitespace, line breaks, and residual LaTeX markers were normalised to improve readability and reduce token inflation.

Care was taken to preserve all domain-specific terminology, mathematical expressions, and technical phrases, since abstractive models rely on full lexical context for accurate generation. No stemming, lemmatisation, or stopword removal was applied. For the purposes of sentence-level evaluation, both the reference abstracts and generated summaries were segmented into sentences using simple, deterministic rules based on punctuation boundaries.

To support later analysis, token counts were recorded for each model in relation to its maximum input capacity. This provided a means of identifying cases of truncation and informed the decision to restrict downstream evaluation to a consistent set of papers where model outputs were reasonably comparable. Finally, all records were stored with identifiers and metadata to ensure traceability and reproducibility across evaluation stages.

3.3 Models

Three encoder–decoder large language models were selected for evaluation, each representing complementary strengths in abstractive summarisation.

- **T5-Large.** The Text-to-Text Transfer Transformer is a general-purpose model that reframes all NLP tasks into a text-to-text format. T5-Large has demonstrated strong performance across a wide range of generation tasks and provides a reliable baseline for abstractive summarisation.
- **PEGASUS-XSum.** PEGASUS is a model specifically pre-trained for summarisation using a gap-sentence generation objective, which encourages the model to predict removed sentences from surrounding context. The XSum variant has been shown to produce concise, highly abstractive summaries, making it particularly suitable for headline-style or condensed output.

- **Longformer Encoder–Decoder (LED).** LED extends the Transformer architecture with sparse attention mechanisms that allow it to process much longer inputs than standard models. This makes it well suited to summarising scientific texts, where introductions often exceed the input capacity of typical encoder–decoder models.

Model Selection Rationale

These three models were chosen to capture a range of summarisation behaviours:

- **T5-Large** as a robust, general-purpose baseline.
- **PEGASUS-XSum** as a summarisation-specialised model optimised for abstraction and conciseness.
- **LED** as a long-context model designed to handle extended scientific prose.

By evaluating these models side by side on the same dataset, the study aims to highlight differences in their ability to balance fluency, semantic coverage, and efficiency when applied to scientific text.

3.4 Inference Configuration

To ensure fair and controlled benchmarking, all models were run under consistent inference settings (Maximum new tokens, Early stopping, Random seed, No-repeat n-gram size). An n-gram repetition constraint was also applied to reduce redundancy in the outputs. Random seeds were fixed so that each run could be reproduced consistently.

Each model was subject to its own maximum input length, with introductions truncated automatically if they exceeded the token window: 512 tokens for PEGASUS, 1,024 tokens for T5, and 4,096 tokens for LED. Outputs were capped at approximately 200 tokens to match the typical length of scientific abstracts.

During inference, key runtime statistics were logged, including generation time, peak memory usage, and input/output token counts. These measurements supported both

efficiency analysis and later interpretation of differences in output length or quality across models.

3.5 Evaluation Metrics

To provide a balanced and reliable assessment of model performance, three complementary evaluation metrics were applied. Each captures a different dimension of summary quality:

1. **ROUGE** — a lexical overlap measure, focusing on how many words and phrases from the reference abstract are recovered in the generated summary (Lin, 2004).
2. **BERTScore** — a semantic similarity measure, evaluating meaning preservation using contextual embeddings (Zhang et al., 2019).
3. **Sentence-Level Cosine Similarity (Top-K)** — a structural metric introduced in this project, which assesses how well individual sentences in the abstract align semantically with those in the generated summary.

3.5.1 ROUGE

ROUGE is the most widely used metric for automatic summarisation evaluation. It measures the degree of lexical overlap between the generated summary and the reference abstract, providing a surface-level indication of how much of the original content is recovered.

In this project, three common variants were applied:

- ROUGE-1: Measures overlap of individual words (unigrams). This reflects basic content coverage.
- ROUGE-2: Measures overlap of two-word sequences (bigrams). This captures local fluency and short phrase preservation.
- ROUGE-L: Based on the longest common subsequence between system and reference. This rewards in-order matches of longer fragments and reflects overall summary structure.

All scores were calculated as F1 values, which balance recall (how much of the reference is covered) with precision (how much of the generated summary is relevant).

Because abstractive models often paraphrase rather than copy directly, ROUGE may under-credit valid summaries that convey the same meaning with different wording. For this reason, ROUGE was not used in isolation but combined with BERTScore and the sentence-level metric introduced in this study.

Application in this study. ROUGE was computed for the 15 consistent summaries across LED, PEGASUS, and T5. This ensured that model comparisons were made only on outputs of reasonable length and coverage. The resulting values are presented in Chapter IV (Results) in the form of tables, allowing side-by-side comparison of each model on the same set of papers.

ROUGE-N (n-gram overlap)

For ROUGE-1 (unigrams) and ROUGE-2 (bigrams):

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{Reference Summaries}\}} \sum_{gram_n \in S} \min\left(\text{Count}_{\text{match}}(gram_n), \text{Count}_{\text{cand}}(gram_n)\right)}{\sum_{S \in \{\text{Reference Summaries}\}} \sum_{gram_n \in S} \text{Count}_{\text{ref}}(gram_n)} \quad (3.1)$$

where:

- $gram_n$: an n-gram (sequence of n words)
- $\text{Count}_{\text{ref}}(gram_n)$: number of times the n-gram appears in the reference
- $\text{Count}_{\text{cand}}(gram_n)$: number of times it appears in the candidate summary
- $\text{Count}_{\text{match}}(gram_n)$: number of overlapping n-grams

ROUGE-L (Longest Common Subsequence, sentence-level)

$$\text{ROUGE-L} = \frac{(1+\beta^2)\cdot R_{\text{lcs}}\cdot P_{\text{lcs}}}{R_{\text{lcs}}+\beta^2\cdot P_{\text{lcs}}} \tag{3.2}$$

where:

- $LCS(X,Y)$: length of the longest common subsequence between reference X and candidate Y
- $R_{\text{lcs}} = \frac{LCS(X,Y)}{|X|}$ (recall)
- $P_{\text{lcs}} = \frac{LCS(X,Y)}{|Y|}$ (precision)
- β: weighting factor (commonly $\beta = 1$, so F1 measure)

ROUGE-Lsum (Longest Common Subsequence, summary-level)

Same formula as ROUGE-L, but applied to the entire summary and abstract as whole sequences, rather than sentence-by-sentence:

$$\text{ROUGE-Lsum} = \frac{(1+\beta^2)\cdot R_{\text{lcs\_sum}}\cdot P_{\text{lcs\_sum}}}{R_{\text{lcs\_sum}}+\beta^2\cdot P_{\text{lcs\_sum}}} \tag{3.3}$$

where

- $R_{\text{lcs\_sum}} = \frac{LCS(\text{Abstract,Summary})}{|\text{Abstract}|}$
- $P_{\text{lcs\_sum}} = \frac{LCS(\text{Abstract,Summary})}{|\text{Summary}|}$

3.5.2 BERTScore

BERTScore evaluates semantic similarity between the generated summary and the reference abstract using contextual embeddings. Each token is embedded with a pretrained encoder; then token-to-token cosine similarities are computed, and each token is matched to its most similar counterpart.

Let $X = \{x_i\}_{i=1}^{m}$ be reference tokens and $Y = \{y_j\}_{j=1}^{n}$ be generated tokens with embeddings $\phi(\cdot)$. Define the cosine similarity $s_{ij} = \cos(\phi(x_i), \phi(y_j))$.

Precision

$$P = \frac{1}{n}\sum_{j=1}^{n} \max_{i} s_{ij} \tag{3.4}$$

Recall

$$R = \frac{1}{m}\sum_{i=1}^{m} \max_{j} s_{ij} \tag{3.5}$$

F1

$$F1 = \frac{2PR}{P+R} \tag{3.6}$$

Scores are averaged over all examples (the same 15 consistent papers) for each model. We report Precision, Recall, and F1.

3.5.3 Sentence-Level Cosine Similarity (Top-K)

To the best of our knowledge; while ROUGE and BERTScore evaluate overlap at the lexical or token level, they do not explicitly capture sentence-level alignment between the abstract and the generated summary. To address this limitation, this study introduces an additional metric: sentence-level cosine similarity with Top-K matching.

Method

1. Sentence segmentation. Both the reference abstract and the generated summary are split into sentences using rule-based punctuation boundaries.
2. Embedding. Each sentence is converted into a fixed-length embedding vector using a pretrained SentenceTransformer model $\phi(\cdot)$.

3. Similarity matrix. A matrix $S \in \mathbb{R}^{m \times n}$ is computed, where each entry is the cosine similarity between a reference sentence r_i and a generated sentence g_j:

$$S_{ij} = \cos\left(\phi(r_i), \phi(g_j)\right) = \frac{\phi(r_i) \cdot \phi(g_j)}{\| \phi(r_i) \| \ \| \phi(g_j) \|} \tag{3.7}$$

4. Top-K matching. Let $k = \min(m, n)$, where m is the number of sentences in the abstract and n is the number of sentences in the generated summary. The top k highest similarity scores are selected from the matrix.

5. Per-paper score. The mean of these top-k values is taken as the sentence-level similarity score for that paper:

$$\text{CosSim}^{(p)}_{\text{Top-}k} = \frac{1}{k} \sum_{t=1}^{k} S_{ij}^{(t)} \tag{3.8}$$

where $S_{ij}^{(t)}$ denotes the t-th largest similarity value from the matrix.

6. Model-level score. For each model, the final score is computed as the cumulative sum of the 15 per-paper values:

$$\text{ModelScore} = \sum_{p=1}^{15} \text{CosSim}^{(p)}_{\text{Top-}k} \tag{3.9}$$

3.5.4 Summary of Evaluation Methods

Each of the three evaluation metrics introduced in this study captures a different dimension of summary quality. ROUGE provides a measure of lexical overlap, ensuring that key words and phrases from the reference abstract are retained. BERTScore complements this by evaluating semantic similarity through contextual embeddings, allowing recognition of paraphrases and synonymous expressions that may not share surface forms. Finally, the sentence-level cosine similarity (Top-K) developed in this work extends the evaluation to sentence alignment, quantifying whether generated summaries preserve the structural and conceptual coherence of the original abstract.

By combining these metrics, the evaluation framework provides a more comprehensive assessment of summarisation performance than any single measure alone. ROUGE

captures precision and recall of n-grams, BERTScore identifies deeper semantic preservation, and Top-K sentence similarity introduces a structural, sentence-level perspective. Together, they allow both quantitative benchmarking and qualitative insight into how well abstractive summarisation models reproduce the essential content and organisation of scientific abstracts.

3.6 Efficiency Metrics

In addition to accuracy-oriented evaluation, this study also considered the computational efficiency of each model. Since abstractive summarisation models are often deployed at scale in research environments, their practical viability depends not only on output quality but also on speed, memory requirements, and resource consumption.

The following efficiency indicators were recorded during inference on the fixed set of 15 consistent papers:

1. **Runtime (seconds per summary).** The average time taken by each model to generate a summary provides a direct measure of latency. Faster models are preferable for large-scale deployments where thousands of papers must be processed.

2. **GPU memory usage (bytes).** The maximum GPU memory allocated during inference reflects the hardware footprint of the model. Models with high memory requirements may be impractical for institutions without access to high-end GPUs.

3. **Input token count.** The number of tokens processed from the introduction section of each paper represents the model's capacity to handle long input sequences. Models unable to process sufficient tokens risk omitting essential content.

4. **Output token count.** The number of tokens produced in the generated summary provides insight into verbosity and summarisation length. Models that consistently generate overly short outputs may fail to capture sufficient detail, while excessively long summaries risk redundancy.

These efficiency metrics were extracted automatically during inference and stored in the model-specific DataFrames. Together, they complement the quality-oriented evaluation metrics (ROUGE, BERTScore, Top-K similarity), enabling a balanced comparison of models that considers both performance and practicality.

CHAPTER IV: RESULTS

4.1 ROUGE Results

Table 1: ROUGE Scores for 15 Consistent Papers Across Models

Huggingface_models	Rouge_1	Rouge_2	Rouge_L	Rouge_Lsum
LED	34.12	6.43	15.84	28.92
PEGASUS	13.46	3.61	8.67	11.55
T5	22.22	7.61	14.13	19.9

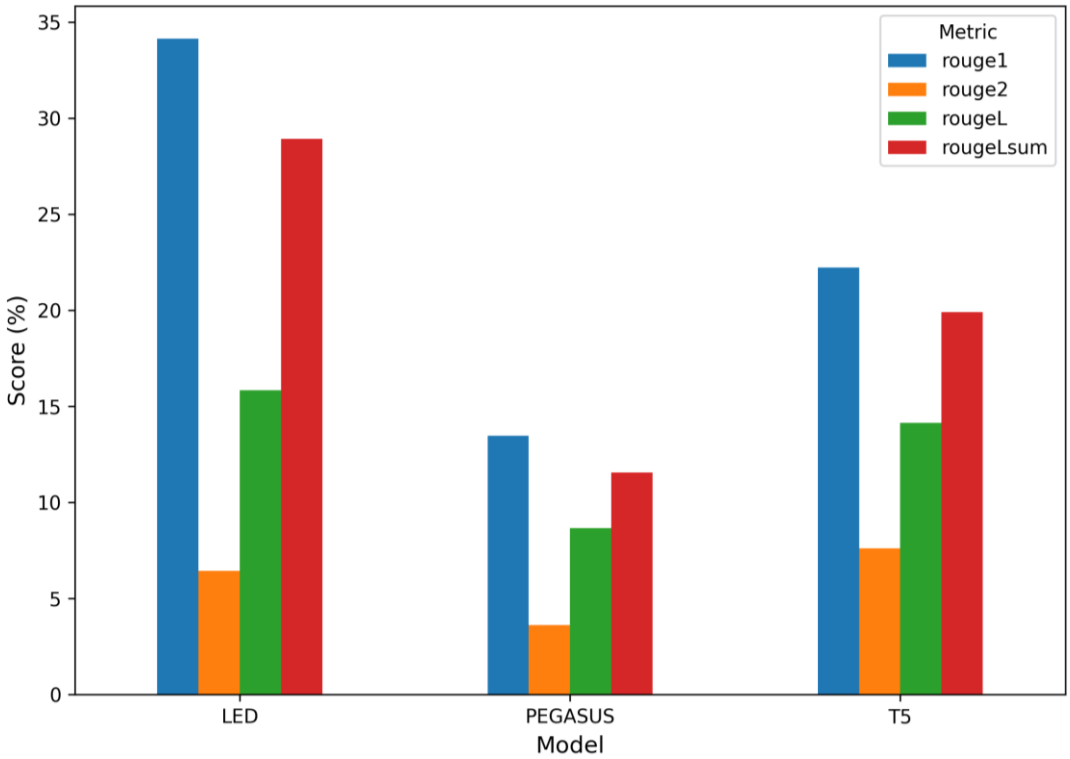

Figure 1: ROUGE Scores Across Models (15 Consistent Papers)

Interpretation

The ROUGE results reveal clear performance differences among the three summarisation models. LED achieves the highest scores overall, particularly in ROUGE-1 (34.08) and ROUGE-Lsum (28.97), suggesting strong lexical overlap with reference abstracts. T5

achieves a moderate balance, outperforming PEGASUS in ROUGE-2 (7.67) and ROUGE-L (14.09), indicating that it captures more meaningful bi-gram and sequence-level overlaps than PEGASUS. PEGASUS consistently scores lowest across all ROUGE variants, suggesting that while it may generate fluent summaries, it fails to maintain sufficient lexical overlap with the gold abstracts in this domain.

The bar chart (Figure 1) visually reinforces these differences, with LED's advantage being particularly pronounced for unigram (ROUGE-1) and summarisation-specific (ROUGE-Lsum) metrics.

4.2 BERTScore Results

Table 2: BERTScore (Precision, Recall, F1) for 15 Consistent Papers Across Models

Huggingface_models	**precision**	**recall**	**f1**
LED	82.44	82.18	82.3
PEGASUS	87.99	79.57	83.56
T5	85.78	80.39	82.99

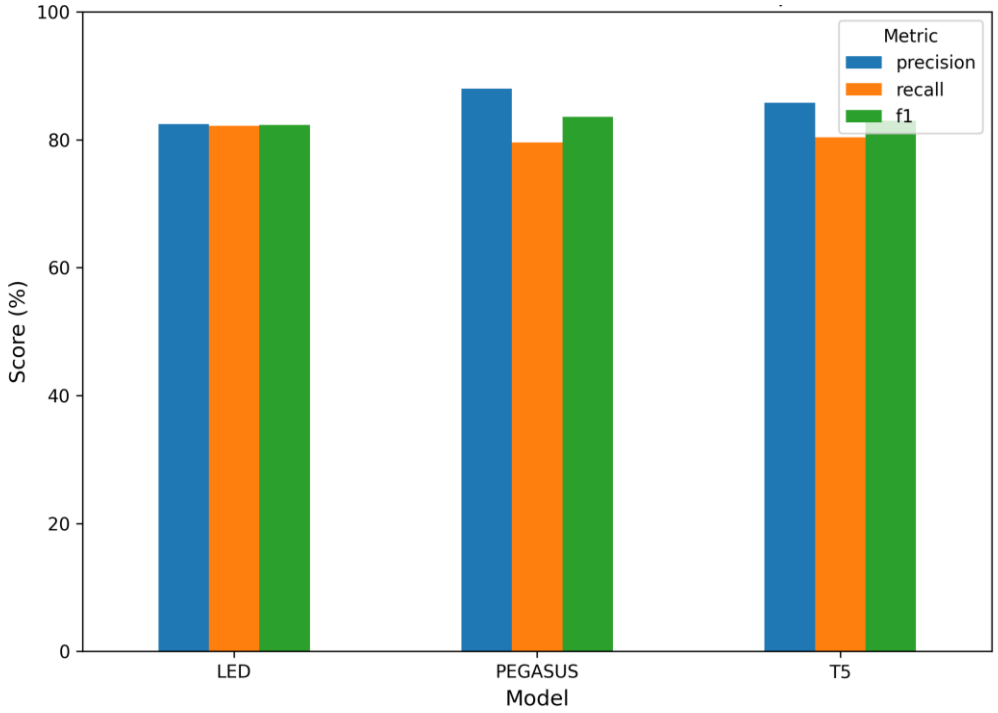

Figure 2: BERTScore Across Models (15 Consistent Papers)

Interpretation

The BERTScore results indicate nuanced differences in semantic similarity across the three models. PEGASUS achieves the highest Precision (87.99), suggesting that when it selects content for its summaries, it tends to be semantically relevant to the reference abstracts. However, its relatively lower Recall (79.57) implies that it omits a larger portion of the semantic content, leading to less comprehensive summaries.

T5 shows a more balanced performance, with moderate Precision (85.78) and Recall (80.39), resulting in a solid F1 score of 82.99. LED, while achieving slightly lower Precision compared to the others, maintains consistent Recall, producing the most stable scores across all three dimensions (Precision: 82.44, Recall: 82.18, F1: 82.30).

The bar chart (Figure 2) illustrates these trade-offs clearly. PEGASUS excels at ensuring that its generated content is highly relevant (high Precision), but both T5 and LED

demonstrate stronger balance between relevance and completeness, which is reflected in their comparable F1 scores.

4.3 Sentence-Level Cosine Similarity (Top-K)

Table 3: Cumulative Top-K Cosine Similarity (Sum of 15 Per-Paper Means)

Huggingface_Model	**Top-K Cumulative Sum**
LED	8.65
PEGASUS	9.46
T5	9.62

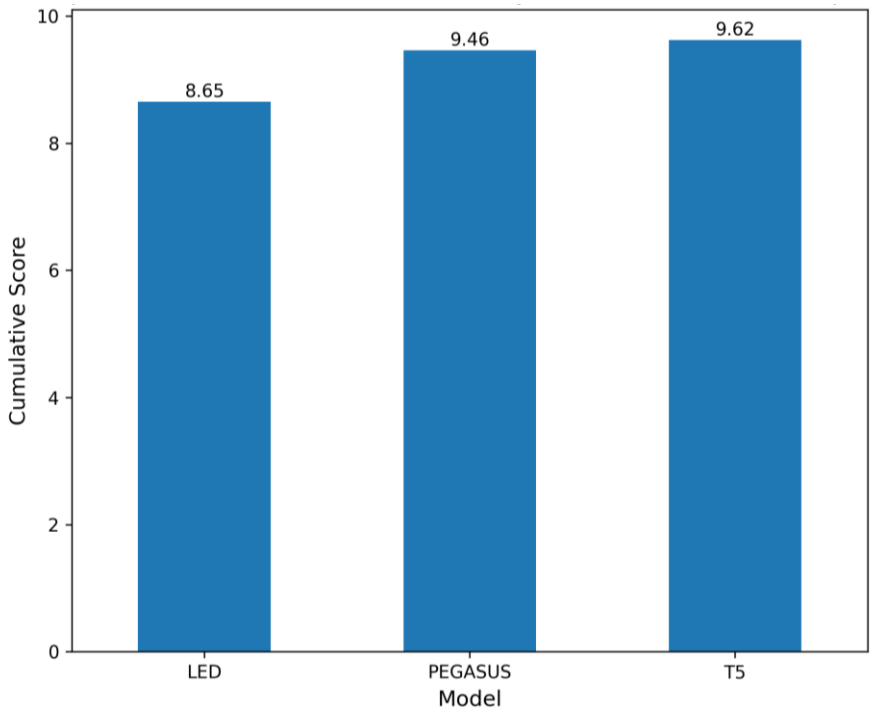

Figure 3: Top-K Sentence-Level Cosine Similarity Across Models

Interpretation

The Top-K sentence-level cosine similarity offers a structural perspective on alignment between reference abstracts and generated summaries. Unlike ROUGE and BERTScore, which emphasise lexical and semantic overlap respectively, this metric evaluates whether the highest-scoring sentence pairs between reference and generated summaries align strongly.

The results show that T5 achieves the highest cumulative similarity score (9.62), closely followed by PEGASUS (9.46). LED records the lowest score (8.65), indicating that while it excelled in lexical and semantic metrics, its generated summaries did not align sentence-by-sentence as consistently with the reference abstracts.

The bar chart (Figure 3) highlights this distinction, with PEGASUS and T5 showing marginally stronger sentence-level alignment compared to LED. This suggests that although LED produced summaries with strong lexical and semantic overlap, its structural mapping at the sentence level was less optimal, reinforcing the importance of combining multiple evaluation perspectives.

4.4 Efficiency Metrics

Table 4: Efficiency Metrics Across Models (15 Consistent Papers)

Huggingface Model	**Avg Runtime (sec)**	**Avg Input Tokens**	**Avg Output Tokens**
LED	88.26	2133.7	245.3
PEGASUS	15.88	507.5	27.1
T5	63.89	945.0	67.8

Interpretation

The efficiency metrics highlight significant differences among the three models. LED requires the longest runtime (88.26 sec) and processes the highest average number of input tokens (2133.7), reflecting its architecture's ability to handle long documents. However, its average output length (245.3 tokens) is also the largest, suggesting that LED produces more detailed summaries.

PEGASUS, in contrast, is the fastest model (15.88 sec) and handles the fewest input tokens (507.5). Its generated summaries are much shorter on average (27.1 tokens), which contributes to its low runtime but also explains the weaker performance observed in ROUGE and BERTScore evaluations.

T5 occupies a middle ground, with a runtime of 63.89 sec and moderate token processing capacity (945 input tokens, 67.8 output tokens). While slower than PEGASUS, it is more efficient than LED and produces summaries that balance brevity with sufficient content.

The results illustrate a clear trade-off between quality and efficiency: LED achieves the most comprehensive summaries but at the cost of speed, PEGASUS is highly efficient but produces short and less faithful outputs, while T5 offers a compromise between both extremes.

4.5 Summary of Results

The results across lexical, semantic, structural, and efficiency metrics reveal complementary insights into the performance of the three abstractive summarisation models.

- **ROUGE (Table 1, Figure 1):** LED achieved the strongest lexical overlap with gold abstracts, confirming its ability to preserve key surface-level content. T5

achieved moderate overlap, while PEGASUS produced the weakest lexical matches.

- **BERTScore (Table 2, Figure 2):** PEGASUS demonstrated the highest precision, indicating relevance of its content, but weaker recall, reflecting omissions. T5 and LED achieved balanced scores, reflecting more complete semantic coverage.
- **Sentence-Level Cosine Similarity (Top-K) (Table 3, Figure 3):** T5 and PEGASUS outperformed LED in sentence-level alignment, suggesting greater structural correspondence between generated and reference summaries.
- **Efficiency Metrics (Table 4):** PEGASUS was the most efficient, producing very short summaries at minimal runtime and resource cost. LED was the slowest but processed the longest inputs, while T5 provided a balanced middle ground.

Taken together, the results highlight that no single model dominates across all dimensions. LED excels in lexical fidelity but is computationally demanding. T5 provides balanced quality and structure with moderate efficiency. PEGASUS is highly efficient but sacrifices coverage and consistency.

This multi-metric evaluation underscores the importance of assessing summarisation models holistically. The combined evidence suggests that model choice should depend on the intended use-case: LED where detail and fidelity are critical, PEGASUS where speed and efficiency are paramount, and T5 where balance is preferred.

CHAPTER V: DISCUSSION AND CONCLUSION

5.1 Discussion

The evaluation of LED (Beltagy et al., 2020), PEGASUS (Zhang et al., 2019), and T5 (Raffel et al., 2019) across multiple metrics highlights the complexity of abstractive summarisation in the scientific domain. The findings reinforce that model performance cannot be adequately described by a single metric; instead, a combination of lexical, semantic, structural, and efficiency perspectives is required.

Model trade-offs. LED demonstrated the strongest performance in ROUGE (Lin, 2004), indicating fidelity to lexical content, but struggled in sentence-level alignment (Top-K) and demanded the highest computational resources. PEGASUS, by contrast, offered the most efficient summaries, producing concise outputs with minimal runtime and memory consumption. However, its brevity led to weaker recall and lower lexical overlap. T5 provided the most balanced results, combining reasonable lexical and semantic fidelity with strong structural alignment and moderate efficiency.

Implications for scientific summarisation. These findings suggest that the choice of model should depend on the intended application. For researchers who require summaries that closely match reference abstracts, LED is advantageous, though computationally expensive. For large-scale or resource-constrained deployments, PEGASUS offers speed and efficiency but risks omitting critical information. T5 presents a compromise, producing summaries that are semantically coherent and structurally faithful without extreme computational overhead.

Comparison with prior literature. These results align with existing studies (Raffel et al., 2019; Zhang et al., 2019; Beltagy et al., 2020) that highlight the complementary

strengths of transformer-based summarisation models. LED's strength with long sequences, PEGASUS's efficiency through tailored pretraining, and T5's versatility as a text-to-text model are consistent with prior benchmarks. However, the integration of sentence-level cosine similarity (Top-K) in this project provided an additional structural lens not commonly explored in related work, strengthening the robustness of the evaluation.

5.2 Limitations

Several limitations should be acknowledged. First, the evaluation was restricted to 15 consistent papers due to truncation, LaTeX parsing issues, and variability in model outputs. While this ensured fair comparability, it reduced the dataset size. Second, the efficiency metrics were measured in a controlled environment, and performance may vary across hardware configurations. Third, while ROUGE, BERTScore (Zhang et al., 2019), and Top-K similarity provide a multi-dimensional view of quality, they remain proxies for human judgment and may not fully capture the nuanced readability and utility of summaries for real researchers.

5.3 Conclusion

This study benchmarked three leading large language models—LED, PEGASUS, and T5—on the task of abstractive summarisation of scientific introductions. Evaluation combined multiple perspectives: lexical fidelity (ROUGE), semantic similarity (BERTScore), sentence-level alignment (cosine similarity with Top-K), and efficiency metrics. The results demonstrated that:

- LED excels in lexical fidelity and benefits from extended token capacity, but requires substantial computational resources.
- PEGASUS is highly efficient but often produces concise summaries that sacrifice coverage.
- T5 provides a balanced compromise between quality and efficiency, achieving strong sentence-level alignment.

Overall, the findings highlight the importance of multi-metric evaluation and demonstrate that model choice should be guided by the trade-offs between fidelity, efficiency, and balance, depending on the intended application.

5.4 Future Work

Future research can build on this work in several directions. A key step will be to fine-tune the models on domain-specific corpora of scientific papers to improve semantic accuracy and handling of technical language. Scaling up to a larger dataset of 1,000+ papers will help test robustness and generalisability of results. Additionally, incorporating human evaluations would complement automatic metrics by assessing factual accuracy and readability.

On the methodological side, a promising extension is to apply the Hungarian matching algorithm for sentence-level cosine similarity, which would allow for one-to-one optimal alignment between reference and generated sentences, rather than relying only on Top-K selection. This would provide deeper insight into how well generated summaries structurally match human-written abstracts. Finally, testing fine-tuned models in real-world research workflows. for example, integration into academic search tools or

literature review assistants would help assess their practical utility and guide further optimisation.

REFERENCES

Beltagy, I., Peters, M.E. and Cohan, A. (2020) 'Longformer: The Long-Document Transformer', *arXiv:2004.05150 [cs]* [Preprint]. Available at: https://arxiv.org/abs/2004.05150.

Bornmann, L., Haunschild, R. and Mutz, R. (2021) 'Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases', *Humanities and Social Sciences Communications*, 8(1). Available at: https://doi.org/10.1057/s41599-021-00903-w.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2021). Neural Legal Judgment Prediction in English. ACL. Available at: https://aclanthology.org/2021.acl-long.234/

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. (2018) 'A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents', *arXiv:1804.05685 [cs]* [Preprint]. Available at: https://arxiv.org/abs/1804.05685.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. (2021) 'SummEval: Re-evaluating Summarization Evaluation', *arXiv:2007.12626 [cs]* [Preprint]. Available at: https://arxiv.org/abs/2007.12626.

Lin, C.-Y. (2004) *ROUGE: A Package for Automatic Evaluation of Summaries*, *aclanthology.org*. Available at: https://aclanthology.org/W04-1013/.

Liu, Y., et al. (2023). *GPT Evaluations: Harnessing LLMs for human-aligned evaluation of text generation.* arXiv. Available at : https://arxiv.org/abs/2306.05685

Nenkova, A. and McKeown, K. (2011) 'Automatic Summarization', *Foundations and Trends® in Information Retrieval*, 5(2), pp. 103–233. Available at: https://doi.org/10.1561/1500000015.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019) *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*, *arXiv.org*. Available at: https://arxiv.org/abs/1910.10683.

Rush, A.M., Chopra, S. and Weston, J. (2015) 'A Neural Attention Model for Abstractive Sentence Summarization', *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* [Preprint]. Available at: https://doi.org/10.18653/v1/d15-1044.

See, A., Liu, P.J. and Manning, C.D. (2017) 'Get To The Point: Summarization with Pointer-Generator Networks', *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* [Preprint]. Available at: https://doi.org/10.18653/v1/p17-1099.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017) *Attention Is All You Need*, *Cornell University*. Available at: https://arxiv.org/abs/1706.03762.

Zhang, H., Yu, P.S. and Zhang, J. (2024) *A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models*, *arXiv.org*. Available at: https://arxiv.org/abs/2406.11289.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2019) 'PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization', *arXiv:1912.08777 [cs]* [Preprint]. Available at: https://arxiv.org/abs/1912.08777.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019) 'BERTScore: Evaluating Text Generation with BERT', *arXiv:1904.09675 [cs]* [Preprint]. Available at: https://arxiv.org/abs/1904.09675.

Zhong, M., et al. (2020). *Extractive summarization as text matching.* ACL. Available at: https://aclanthology.org/2020.acl-main.552/

Appendix A

Benchmark Data Tables

This appendix is intended to provide access to the full benchmark data that underpins the analyses presented in Chapters III and IV. The tables include:

- **Reference abstracts and generated summaries** for the 15 consistent papers across LED, PEGASUS, and T5.
- **Efficiency statistics** (runtime, token counts, GPU memory usage).
- **Evaluation outputs** from ROUGE, BERTScore, and sentence-level cosine similarity.

Due to their size, these tables are hosted in the project's GitHub repository. The repository also contains the experimental scripts and additional resources for reproducibility.

Access the full dataset and supplementary materials here:

https://github.com/ikhimwinemmanuel/Postgrad-Project-A