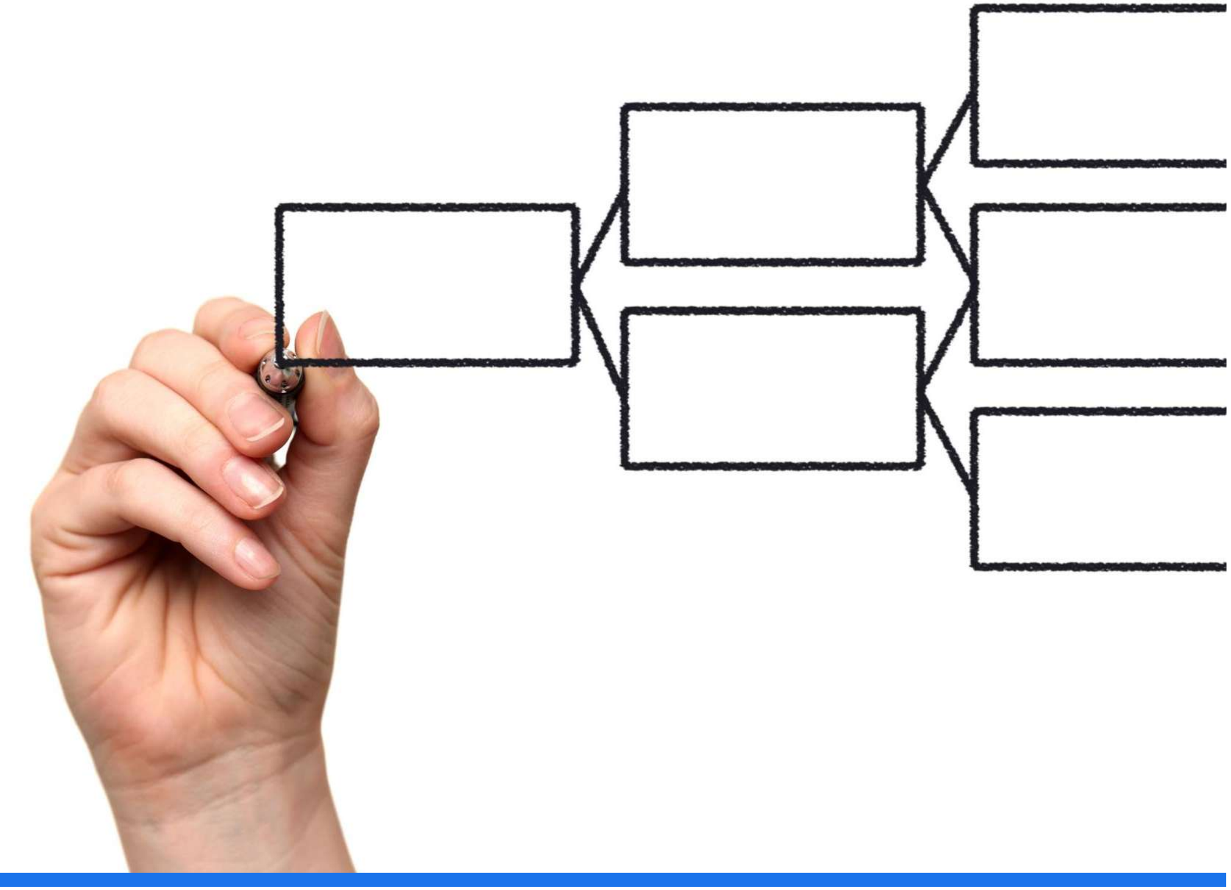
Enhancing Text Summarization Using Flan-T5 Small LLM

EXPLORING ADVANCEMENTS IN SUMMARIZATION TECHNIQUES

Presented By: Nitin Jain



Agenda Items

- Introduction to Text Summarization
- Literature Review
- Understanding Flan-T5 Small Language Model
- Problem Statement
- Experimental Setup
- Methodology
- Results and Discussion
- Conclusion and Future Work

Introduction to Text Summarization

- Automatic text summarisation is the process of extracting main chunks from text documents.
- In this process machine understands the semantic and lexical relation of different words in the Text (Widyassari et al., 2022).
- A well-written text summary should include readability, coherency, syntax, sentence ordering, information diversity, and non-redundancy (Gambhir and Gupta, 2017).

Types of Text Summarization

Extractive Summarization

- Extractive summarization identifies and selects key sentences from the original text, maintaining the original wording and structure.

Abstractive Summarization

- Abstractive summarization involves generating new sentences that represent the main ideas of the text, offering a more concise summary.

Use Cases in NLP

- Both extractive and abstractive summarization methods are valuable in various natural language processing applications, such as content summarization and information retrieval.

Importance of Accurate and Efficient Summarization

Role in Information Retrieval

- Accurate summarization helps readers quickly understand the main ideas of lengthy texts, enhancing their information retrieval process.

Efficiency in Summarization

- Efficient summarization allows for rapid content processing, which is essential in today's fast-paced information environment.

Applications in News and Curation

- Efficient summarization is vital for news aggregation and content curation, facilitating timely updates for users.

Challenges in Current Summarization Methods

Maintaining Coherence

- One challenge is ensuring that generated summaries maintain coherence, making them understandable and logically connected.

Understanding Context

- Summarization methods often struggle with understanding the context, which is essential for accurate representations of the source material.

Reflecting Tone and Intent

- Another challenge is generating summaries that accurately reflect the original tone and intent of the source content.

Literature Review

- Recent research has introduced reinforcement learning (RL) and human feedback loops to improve summary quality, factual consistency, and coherence (Zhong et al., 2022).
- Despite these advances, abstractive text summarization (ATS) remains challenging, particularly due to hallucinations — where models generate plausible but incorrect or unsupported information (Siontis et al., 2024).
- In domain-specific contexts like medical reporting, models have been shown to generate irrelevant or inaccurate details, raising concerns about their reliability. To address such challenges, fine-tuning base models through supervised learning has been employed to better align model knowledge with targeted tasks, such as clinical trial reporting (Markey et al., 2024)

Problem Statement

Why Summarization Matters?

- Summarization condenses essential information for domains like news, law, finance, and medicine.
- Despite advances (e.g., T5, FLAN-T5), current models face critical limitations.

Key Challenges

1. Factual Inconsistencies
   - AI summaries may contain inaccuracies, misrepresentations, or incomplete details.
   - Undermines trust, especially in sensitive domains.
2. Lack of Cohesiveness
   - Summaries may lack logical flow, making them hard to follow or disconnected.

This work explores techniques to improve factual consistency and coherence in AI-generated summaries aiming to enhance their reliability and trustworthiness across critical applications.

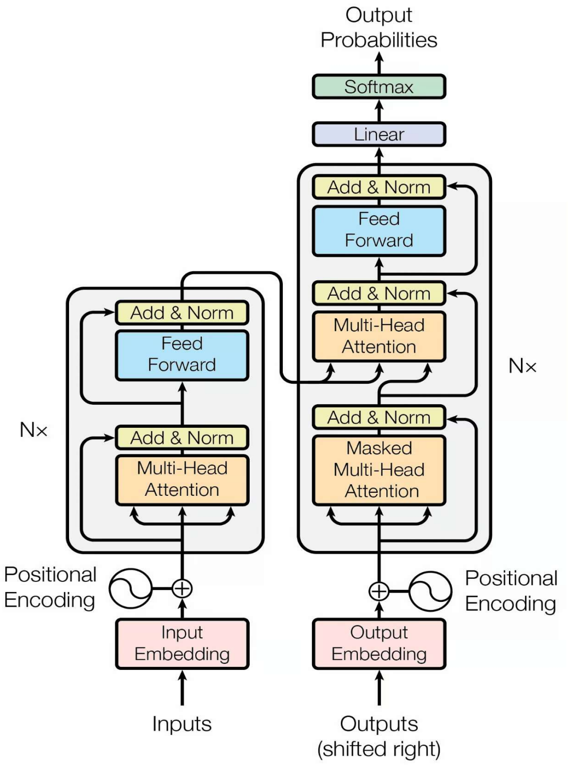
Introduction to Flan-T5 Architecture

- Flan T5 is a transformer-based language model developed by Google, built upon the T5 architecture.

- It's an encoder-decoder model with 12 transformer layers, a feed-forward neural network, and pre-trained on a massive dataset, including web pages, books, and articles.

- The model is designed for various NLP tasks, including text classification, summarization, and question-answering, and is also multilingual.

- Flan T5-Small model used in this study has trained on 80 million model parameters.



Transformer Model Architect (Source)

Key Features and Capabilities

Multi-task Learning

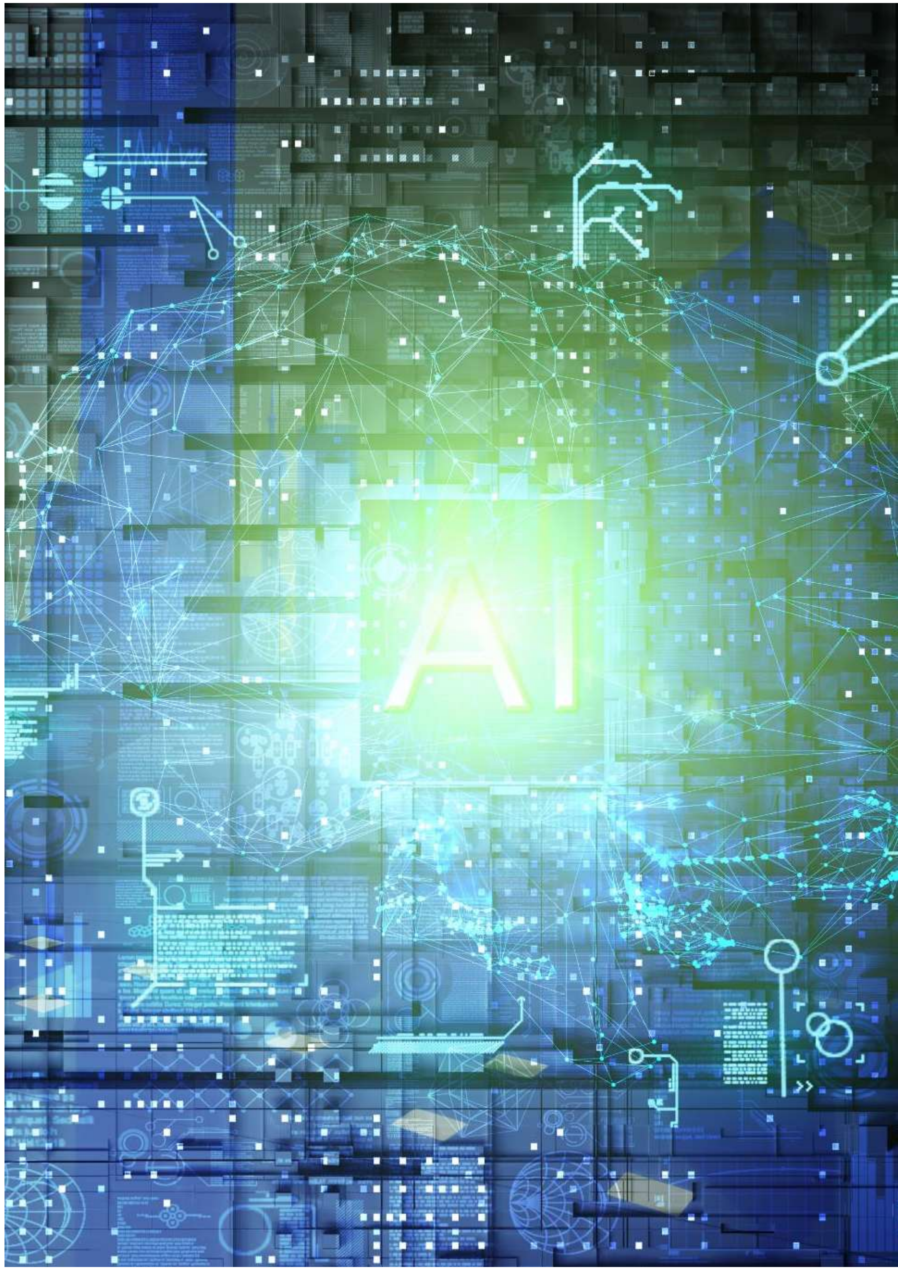
Flan-T5 utilizes multi-task learning to improve its ability to perform various language tasks effectively.

Fine-tuning Capabilities

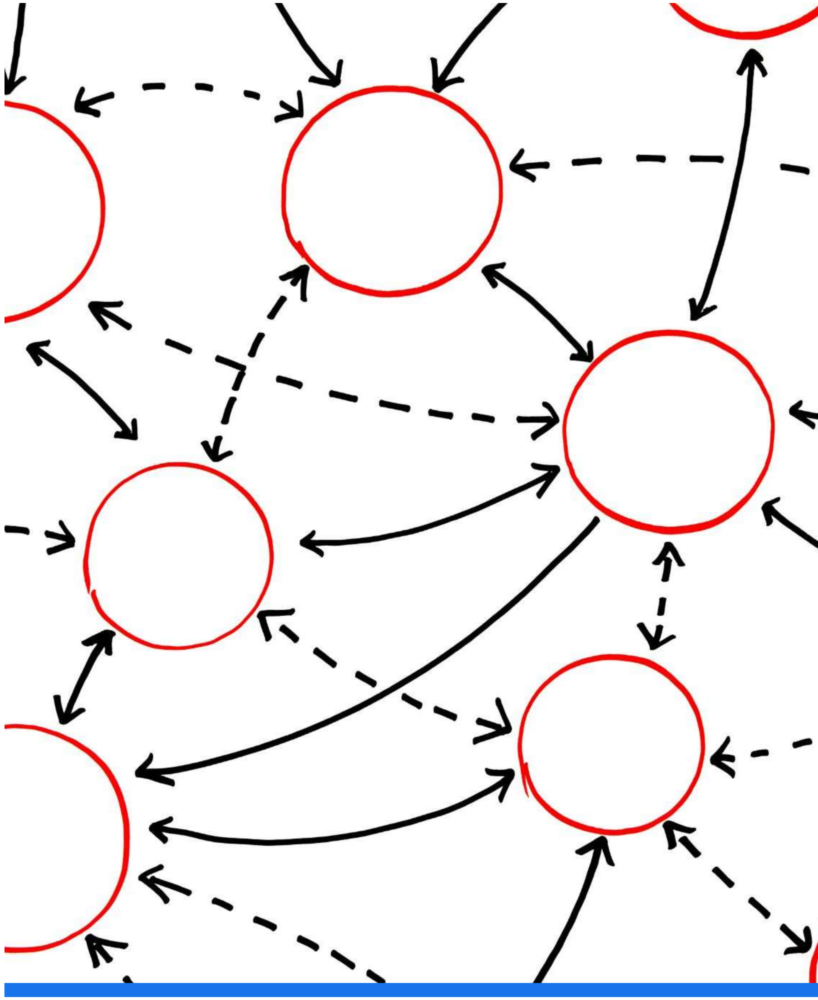
The model's fine-tuning capabilities allow for better adaptation to specific tasks, enhancing overall performance.

Text Summarization Performance

Flan-T5 excels in text summarization, making it a valuable tool for extracting key information from large texts.



Advantages of Flan-T5 for Text Summarization



Improved Context Understanding

Flan-T5 demonstrates enhanced capabilities in comprehending context, leading to better text interpretation during summarization.

Coherent Summaries

The model generates coherent and logically structured summaries, ensuring the essence of the original text is preserved.

Efficiency with Varying Text Lengths

Flan-T5 efficiently processes texts of different lengths, making it versatile for various summarization tasks.

Dataset

- The CNN/Daily Mail dataset is used for this study. This dataset contains news articles paired with human-written summaries and is widely adopted for benchmarking summarisation models.

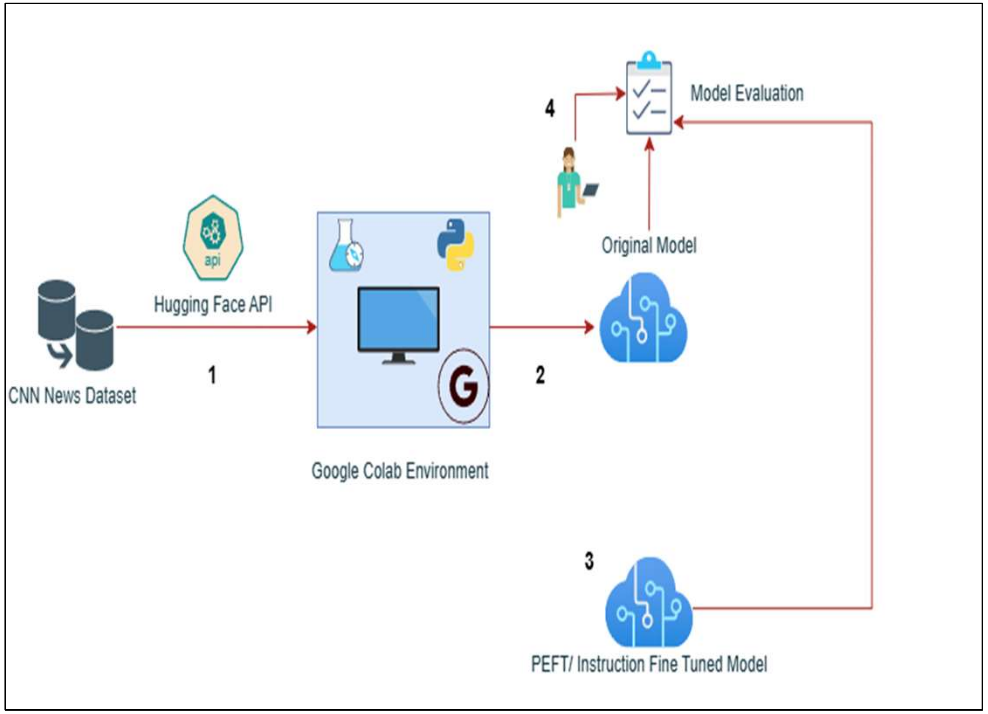
- The dataset taken for this implementation includes 50000 training samples, 5000 validation samples and 2000 testing samples to test the model's behaviour after the training. These testing datasets are shuffled and randomly selected. Each training, validation and testing sample has a news article id, the news article and its highlights.

Experiment Set up

The initial step involves sourcing the Flan T5 – Small model and dataset to the Google Colab environment to ensure they are correctly loaded.

The next step involves fine-tuning a pre-trained large language model using a supervised fine-tuning methodology. Additionally, Parameter Efficient Fine Tuning (PEFT) is applied.

The research methodology then encompasses model evaluation, which will compare the summary generated by the model's vs the reference summary.

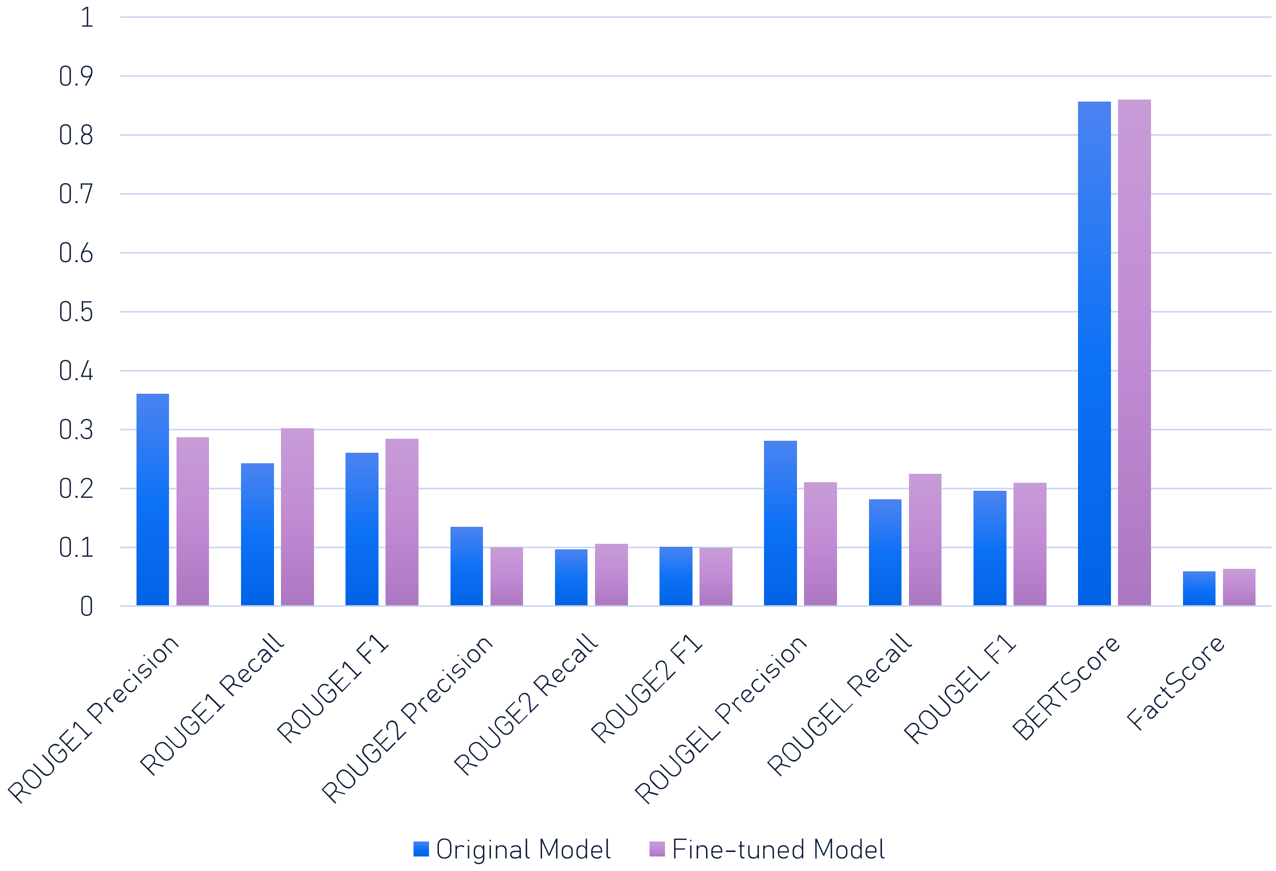


Research Methodology

- This research improves abstractive text summarisation using the Flan-T5 Small model, fine-tuned supervised and PEFT fine-tuning. The methodology includes dataset preparation, supervised fine-tuning, and parameter-efficient fine-tuning using LoRA.
- The dataset is pre-processed to truncate long inputs and structure them in a prompt-based format
- The model training process begins with a task-specific prompt (*"Give the summary of the article: {article}"*), followed by the article text, with the target being the human-written summary of the Flan-T5-Small model
- In the testing phase, the model generated summaries for 2000 test articles without referring to the human-written versions or reference summaries
- Model is then evaluated based on different metrics such as ROUGE-1, ROUGE-L, BERTScore, FactScore.

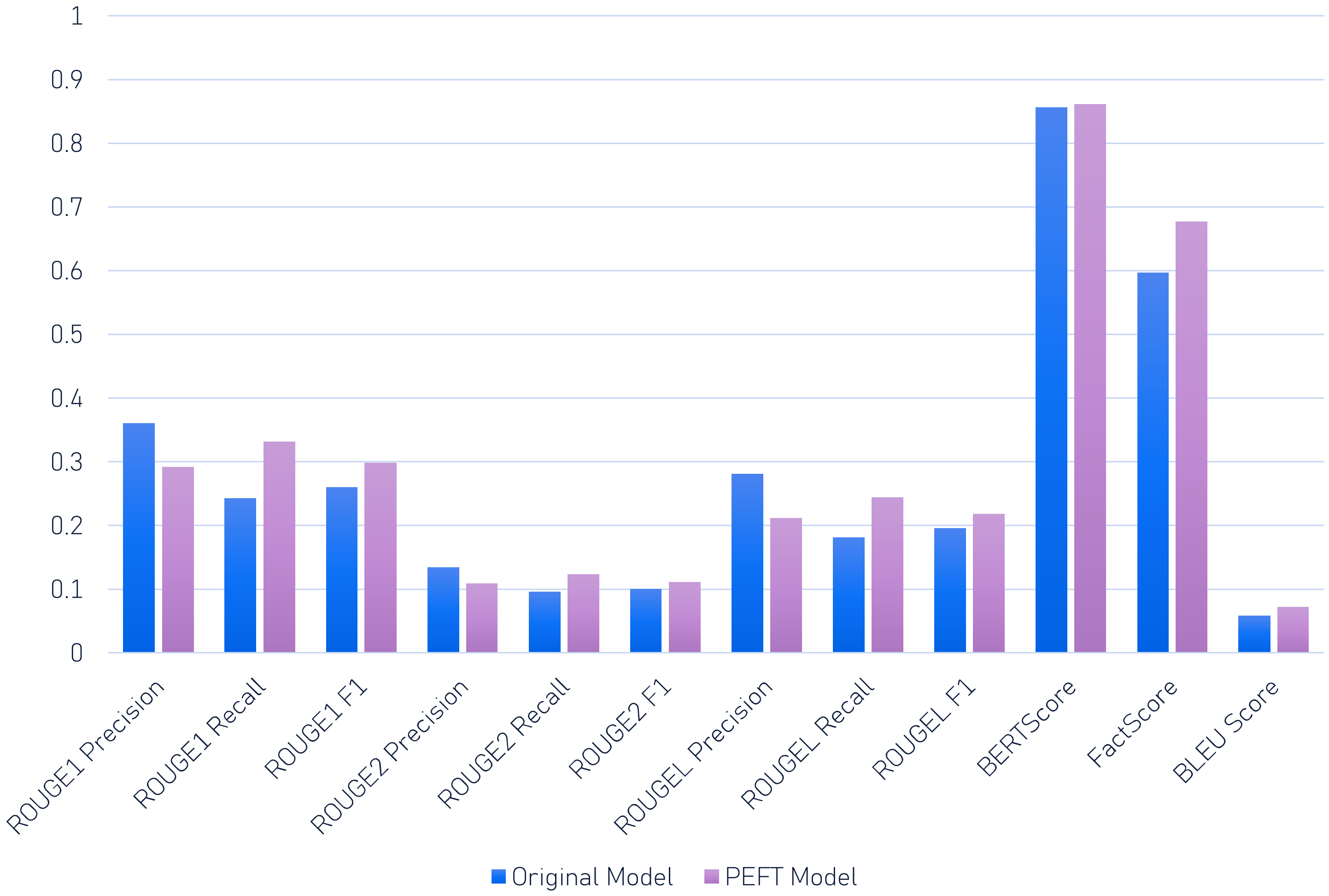
Results and Discussion

Performance Comparison of Fine-Tuned Model with Original Model



- Fine-tuning the model led to notable improvements in recall-oriented and factual metrics, while precision dropped across most ROUGE variants, suggesting a trade-off between content coverage and conciseness.
- ROUGE-1 & ROUGE-L Recall improved significantly (+19.66% and +19.27%), indicating that the fine-tuned model includes more relevant content from the reference summaries. ROUGE-1 and ROUGE-L F1 increased moderately (+8.22% and +6.38%), reflecting overall better balance.
- ROUGE-2 Recall had a small gain (+8.80%), but F1 slightly dropped (-1.45%) due to a sharp decline in precision (-34.36%). All precision values across ROUGE types decreased, especially ROUGE-2 and ROUGE-L, showing that the fine-tuned model tends to generate longer or less precise outputs.
- BERTScore improved slightly (+0.45%), suggesting that the semantic similarity between model-generated and reference summaries improved.
- FactScore increased by +7.82%, highlighting better factual accuracy in generated content.
- BLEU Score improved modestly by +7.08%, reinforcing that the model outputs are now more aligned with reference summaries regarding n-gram overlap.

Performance Comparison of PEFT Model with Original Model



- ROUGE-1 Recall improved by a substantial +36.62%, showing the model captures more relevant information. ROUGE-1 F1 increased by +14.61%, while precision decreased by -19.06%, reflecting the PEFT model's recall-heavy nature. ROUGE-2 Recall jumped +28.93%, while F1 also improved +10.99% despite a -18.91% drop in precision. ROUGE-L Recall improved significantly (+34.60%), with F1 up +11.29% and precision down -24.48%.
- BERTScore saw a modest improvement (+0.62%), indicating slightly better semantic similarity. FactScore increased significantly by +13.43%, showing stronger factual consistency in summaries. BLEU Score rose by +23.51%, showing better n-gram overlap with reference summaries and improved fluency and sentence structure.

Conclusion and Future Work

Summary of Findings

ROUGE Recall scores (ROUGE-1, ROUGE-2, ROUGE-L) showed steady and substantial gains at each stage, with PEFT yielding the highest improvements (e.g., ROUGE-1 Recall jumped from 19.66% to 36.61%.

ROUGE F1 scores also improved consistently, indicating a better balance between recall and precision, though precision decreased slightly after fine-tuning and PEFT.

BERTScore, which measures semantic similarity, increased modestly with fine-tuning (+0.45%) and PEFT (+0.62%), reflecting better meaning preservation in summaries.

FactScore, a critical metric for factual accuracy, improved by 7.82% with supervised fine-tuning and 13.43%, highlighting PEFT's effectiveness in improving factual grounding.

The BLEU Score, which captures n-gram precision and fluency, rose notably from 7.08% to 23.51%, suggesting better syntactic alignment and fluency after PEFT.

Conclusion

Our experiments using the CNN/Daily Mail dataset showcased consistent performance gains post PEFT, with up to 36% improvement in ROUGE Recall and a 13% increase in factual accuracy, confirming the effectiveness of our approach.

Furthermore, the significant improvement in these metrics asserts the better performance for the naïve and small model having fewer parameters, like Flan T5-small.

Moreover, Qualitative analyses indicated better coherence and reduced hallucinations in the generated summaries.

Therefore, this study concludes by providing all the required answers to the research questions. Lastly, further improvements and suggestions are recommended for future work.

Future Work

Multi-Objective Reward Functions

- Incorporating multiple reward components (e.g., factual consistency, readability, novelty) using a weighted sum or Pareto optimisation could lead to more balanced summaries.

Human Feedback Integration

- Incorporating real human preferences or annotator rankings can further align the model with subjective quality factors like tone, coverage, and informativeness.

Robustness and Bias Evaluation

- Future work should evaluate the model's robustness to adversarial prompts and analyse any biases introduced during RLHF to ensure safe deployment.

Thank You!