Improving Educational Text Summarization through Prompt Engineering with FLAN-T5

Mariam Elrafei
Omnia Adel
AlAlamein International University

Spring 2025

Abstract

Text summarization is critical for educational applications, enabling efficient comprehension of complex materials. This study investigates the enhancement of summarization quality using prompt engineering with FLAN-T5, comparing its performance against BART, T5, and TextRank. By designing targeted prompts, we aim to produce concise and accurate summaries tailored for educational contexts. Evaluation using ROUGE metrics on a dataset of 50 educational texts demonstrates that FLAN-T5 with prompt engineering outperforms baseline models, achieving higher scores in ROUGE-1, ROUGE-2, and ROUGE-L. These findings suggest that prompt engineering is a cost-effective approach to improve summarization without extensive model retraining.

1 Introduction

The rapid growth of educational content necessitates tools that can distill complex information into concise summaries for students and educators. Automatic text summarization, both extractive and abstractive, has been widely explored to address this need. Transformer-based models like BART and T5 offer robust abstractive summarization, while TextRank provides a lightweight extractive alternative. Recent advances in prompt engineering allow fine-tuning of model outputs through carefully crafted inputs, avoiding resource-intensive fine-tuning. This paper evaluates the efficacy of prompt engineering with FLAN-T5 for summarizing educational texts, comparing it against established models to assess improvements in summary quality and relevance.

2 Related Work

Text summarization research is divided into extractive methods, which select key sentences, and abstractive methods, which generate new text. BART combines denoising and sequence-to-sequence pretraining, excelling in abstractive tasks. T5 frames summarization as a text-to-text task, offering flexibility across datasets. TextRank, a graph-based algorithm, ranks sentences for extractive summarization. Prompt engineering has emerged as a technique to guide large language models without retraining, showing promise in tasks like summarization. Our work builds on these foundations, focusing on prompt-enhanced FLAN-T5 for educational content.

3 Methodology

We evaluated four summarization models on a dataset of 50 educational texts from the CNN/DailyMail dataset, adapted for educational contexts. The models include:

- BART (`facebook/bart-large-cnn`)
- T5 (`t5-base`)
- TextRank (implemented via Sumy)
- FLAN-T5 (`google/flan-t5-base`) with prompt engineering

For FLAN-T5, we crafted prompts such as `"Generate a concise summary of the following educational text, focusing on key concepts for student understanding."` Each model generated summaries, which were evaluated against human-written reference summaries using ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-L).

4 System Implementation

The models were implemented using Hugging Face Transformers and PyTorch. Text preprocessing was handled with NLTK and SpaCy, while Pandas and Matplotlib facilitated data analysis and visualization. The `rouge-score` library computed ROUGE metrics. The pipeline included text extraction, summary generation, metric computation, and result storage in a CSV file for analysis.

5 Experiments and Results

We measured the performance of each model using ROUGE F1 scores, which assess unigram (ROUGE-1), bigram (ROUGE-2), and longest common subsequence (ROUGE-L) overlap with reference summaries. Table 1 presents the average scores across the 50 samples.

Table 1: Average ROUGE F1 scores for summarization models on 50 educational texts.

Model	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
BART	0.4423	0.2107	0.3856
T5	0.4289	0.1973	0.3712
TextRank	0.4051	0.1754	0.3508
FLAN-T5	0.4672	0.2258	0.4031

FLAN-T5 with prompt engineering consistently outperformed other models, with improvements of 5.7% in ROUGE-1, 7.2% in ROUGE-2, and 4.5% in ROUGE-L over BART. Qualitative analysis showed FLAN-T5 summaries were more coherent and better captured key educational concepts. Example summaries are provided in the appendix.

6 Discussion

The superior performance of prompt-engineered FLAN-T5 highlights the impact of precise input design. By tailoring prompts to emphasize educational relevance, we improved summary qual-

ity without additional computational costs. This approach is particularly valuable in resource-constrained educational settings, where fine-tuning large models is impractical. However, prompt design requires domain expertise, and optimal prompts may vary across contexts.

7 Conclusion

This study demonstrates that prompt engineering with FLAN-T5 significantly enhances text summarization for educational content. By achieving higher ROUGE scores and producing more coherent summaries, FLAN-T5 offers a practical solution for educational applications. Future work could explore automated prompt optimization and broader datasets to further refine this approach.

Appendix

Example summaries and additional qualitative comparisons are available upon request.