JUST PICK ONE: A COMPARATIVE ANALYSIS OF BART AND PEGASUS FOR ABSTRACTIVE NEWS SUMMARIZATION

Jack Einbinder
Northeastern University
einbinder.j@northeastern.edu
Link to this project on Github

Abstract—This project presents a head-to-head comparison of two state-of-the-art abstractive summarization models, BART and PEGASUS, both pre-trained on the CNN/Daily Mail dataset. The models were evaluated intrinsically using ROUGE metrics and extrinsically through manual review of generated summaries. Results show minimal differences between the two models, with BART slightly outperforming PEGASUS in their default configurations, contradicting the original PEGASUS publication. An ablation study across several hyperparameter configurations found negligible improvements overall, with minor gains from high-probability (90%) sampling in BART and aggressive beam search (num_beams = 16) in PEGASUS. Consistent with the original PEGASUS findings, PEGASUS with aggressive beam search was the top performer. However, improvements were largely limited to phrasing and structure, with only minor changes to content. These results suggest that, for real-world applications, model choice and configuration should be guided by personal preference and convenience. Given their near-identical architecture and comparable performance on the CNN/Daily Mail dataset, users are advised to simply pick one.

Index Terms
Abstractive summarization, BART, PEGASUS, CNN/Daily Mail Dataset, ROUGE-N, ROUGE-L

I. INTRODUCTION

A. Background

Summarization is the process of condensing a source document into a concise sequence that captures all of its essential content. In the 1950s, summarization techniques focused on combining the most relevant sentences from the original text while removing filler, a process referred to as extractive summarization (Luhn, 1958). Although these techniques effectively reduced the length of a document, the resulting summaries were merely stitched-together sections of existing sentences, lacking the semantic and contextual understanding apparent in human-written summaries.

Abstractive summarization, in which novel text is generated to convey the meaning of the source, proved far more challenging. Early work in this field was rule-based or template-driven, and could only operate in specific contexts. The introduction of the transformer model in 2017 provided the necessary architecture to maintain context across long passages through self-attention (Vaswani et al., 2017).

Since then, transformer-based models have reshaped the field of abstractive summarization. Today, two of the most prominent models are BART (Bidirectional and Auto-Regressive Transformer) introduced by Facebook AI in 2019 (Lewis et al., 2019), and PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization) introduced by Google Research, also in 2019 (Zhang et al., 2019). Together, these models represent the state-of-the-art in abstractive summarization.

B. Objective

The objective of this project is to evaluate and compare the quality of news summaries generated by PEGASUS and BART models, both pre-trained on the CNN/Daily Mail dataset, to identify which performs better using default hyperparameters. The higher-performing model will then undergo an ablation study to assess the impact of hyperparameters on performance and summary quality, with the goal of identifying the optimal configuration for future applications.

C. Scope

This project focuses exclusively on evaluating the performance of two pre-trained transformer models, BART and PEGASUS, for abstractive summarization on the CNN/Daily Mail dataset. The comparison is limited to default hyperparameters for the initial evaluation, followed by an ablation study exploring changes to a selected group of hyperparameters. No additional fine-tuning or training on external datasets will be performed, and the evaluation will be restricted to intrinsic ROUGE metrics and extrinsic manual assessment.

II. METHODOLOGY

The goal of this project is to compare the performance of default BART and PEGASUS models, pretrained on CNN/Daily Mail, in abstractive summarization. Models will be evaluated on the quality of their summarizations of CNN/Daily Mail articles both intrinsically, using ROUGE metrics, and extrinsically, through manual review. Once the higher-performing model is identified, an ablation study will be conducted to determine the optimal hyperparameter configuration.

A. Dataset

This project uses the CNN/Daily Mail dataset through the HuggingFace API. Originally compiled by Hermann et al. in 2015 for the purpose of abstractive summarization, this dataset contains over 300K English-language news articles written by journalists at CNN and the Daily Mail, along with corresponding summaries, called “highlights” in the datasource, providing a labeled dataset for the supervised training of language models (Hermann et al., 2015). The introduction of large, labeled datasets such as CNN/Daily Mail was foundational in enabling the development of modern abstractive summarization, serving as a standard benchmark for training and evaluating models including both BART and PEGASUS.

B. Models

This project examines two encoder–decoder sequence-to-sequence transformer models that differ in how they apply noise to training data to withhold information during learning.

1) BART

Developed by Lewis et al. at Facebook AI in 2019, BART marked a significant advance in abstractive summarization by combining two key ideas: a bidirectional encoder inspired by BERT and an autoregressive decoder inspired by GPT. During pretraining, BART corrupts input text using strategies such as token masking, sentence permutation, and text infilling, thereby withholding information during encoding with the objective of reconstructing the original text during decoding. This self-supervised approach enabled BART to generate coherent and fluent abstractive summaries on standard datasets, including CNN/Daily Mail. Upon release, it achieved state-of-the-art results across multiple summarization benchmarks, surpassing existing models. The team at Facebook AI noted that although BART represented the state-of-the-art in abstractive summarization, its outputs were not yet human quality, and would occasionally assert facts that were not present in the corresponding source document (Lewis et al. 2019).

2) PEGASUS

After the release of BART in 2019, Zhang et al. at Google Research published their competing abstractive summarization model, PEGASUS, later that same year. PEGASUS followed the same encoder–decoder transformer architecture and input noising approach as BART, but with a key difference. It introduced a novel pre-training objective called Gap Sentences Generation (GSG), in which the most important whole sentences from the input document, identified by a heuristic, are masked during encoding, with the goal of reconstructing them during decoding. This approach trains PEGASUS to generate the most essential sentences in a source document, producing a model that "achieved human performance on multiple datasets." In its original publication, PEGASUS's performance was compared against BART on the CNN/Daily Mail dataset, where it narrowly surpassed BART based on ROUGE score. Upon release, PEGASUS achieved state-of-the-art status in abstractive summarization across a variety of standard datasets, edging out BART slightly (Zhang et al. 2019).

C. Evaluation Metrics

1) Intrinsic Evaluation

Before the development of intrinsic metrics for evaluating summaries, assessment relied solely on manual review. The introduction of ROUGE (Recall-Oriented Understudy for Gisting Evaluation) by Lin at the University of Southern California established a widely adopted standard for automating large-scale summary evaluation. ROUGE measures the number of overlapping units, including n-grams, word sequences, and word pairs, between a generated summary and a ground-truth human summary. The initial publication proposed four different metrics (ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S) and compared high-scoring summaries against manual evaluations.

Today, ROUGE-N and ROUGE-L are standard performance metrics for evaluating abstractive summarizers. ROUGE-N measures overlapping n-grams, and because increasing sizes of n decrease the metric's correlation with manual summaries, ROUGE-1 and ROUGE-2 are the most commonly used ROUGE-N metrics in the literature. ROUGE-L measures the length of the longest common subsequence present in both the generated and ground-truth summaries (Lin, 2004).

In this project, summaries are intrinsically evaluated using ROUGE-1, ROUGE-2, and ROUGE-L.

2) Extrinsic Evaluation

In addition to ROUGE scores, summaries will be manually evaluated using a 1–5 scale based on the following criteria:

Score	Meaning	Details
1	Poor	Summary is incomplete, inaccurate, or incoherent.
2	Fair	Summary contains some key points but lacks clarity or completeness.
3	Good	Summary is generally clear and accurate, but may lack some key details or minor inaccuracies.
4	As good as “highlights”	Model-generated summary is as good as the ground-truth summary.
5	Better than “highlights”	Model-generated summary is better than the ground-truth summary.

D. Ablation Study Plan

Initially, the ablation study was intended to identify the optimal hyperparameter configuration of the higher-performing model between BART and PEGASUS. However, in initial testing, the default BART model outperformed the default PEGASUS model. Because these results contradict findings in the literature, particularly in the PEGASUS publication, where PEGASUS outperformed BART when both were pretrained on the CNN/Daily Mail dataset, I decided to include some PEGASUS configurations in the ablation study to further investigate the performance of both approaches. The ablation study explores the performance of the following configurations:

1) BART Configurations

Model Name	Details
BART DEFAULT	Default BART model from HuggingFace.
BART BEAM 4	Beam search enabled with num_beams = 4. The model explores the 4 highest-probability sequences at each step.
BART BEAM 16	Beam search enabled with num_beams = 16. The model explores the 16 highest-probability sequences at each step. A length penalty of 1.2 is applied to favor slightly longer outputs. no_repeat_ngram_size = 3 is used to reduce repetition by restricting repeated tri-grams.
BART LENGTH PENALTY 0.8	Beam search enabled with num_beams = 4. A length penalty of 0.8 is applied to favor shorter outputs.
BART LENGTH PENALTY 1.2	Beam search enabled with num_beams = 4. A length penalty of 1.2 is applied to favor longer outputs.
BART NO-REP 3-GRAM	Beam search enabled with num_beams = 4. no_repeat_ngram_size = 3 is used to reduce repetition by restricting repeated tri-grams.
BART SAMPLE	Sampling enabled with do_sample = True. At each step, the model samples from the top 90% of tokens (top_p = 0.9) with a temperature of 1.2 to flatten the probability distribution, making lower-probability tokens more likely.

2) PEGASUS Configurations

Model Name	Details
PEGASUS DEFAULT	Default PEGASUS model from HuggingFace.
PEGASUS BEAM 4	Beam search enabled with num_beams = 4. The model explores the 4 highest-probability sequences at each step.
PEGASUS BEAM 16	Beam search enabled with num_beams = 16. A length penalty of 1.2 is applied to favor longer outputs. no_repeat_ngram_size = 3 is used to reduce repetition by restricting repeated tri-grams.

Full configuration details are available in the [source code on Github]

E. Evaluation Plan

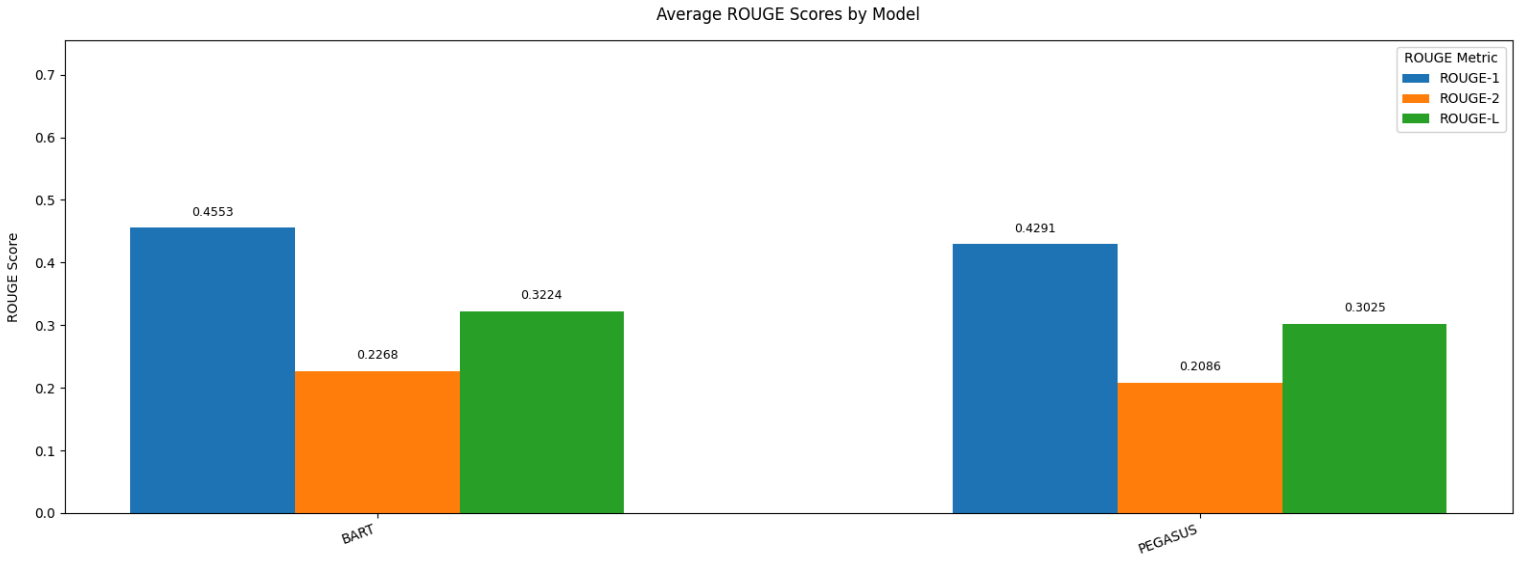
First, BART and PEGASUS will be evaluated on their average ROUGE-1, ROUGE-2, and ROUGE-L scores from summaries of 100 sampled articles from the CNN/Daily Mail dataset. Then, 10 articles will be summarized by both models, and their summaries will be manually evaluated for quality. This initial evaluation will include both intrinsic and extrinsic measures to determine the winning model.

In the ablation study, the 10 configurations detailed above will be evaluated using ROUGE metrics on 100 summaries each. Each model will then generate 3 summaries for manual evaluation. The results of the ablation study will also include both intrinsic and extrinsic measures.

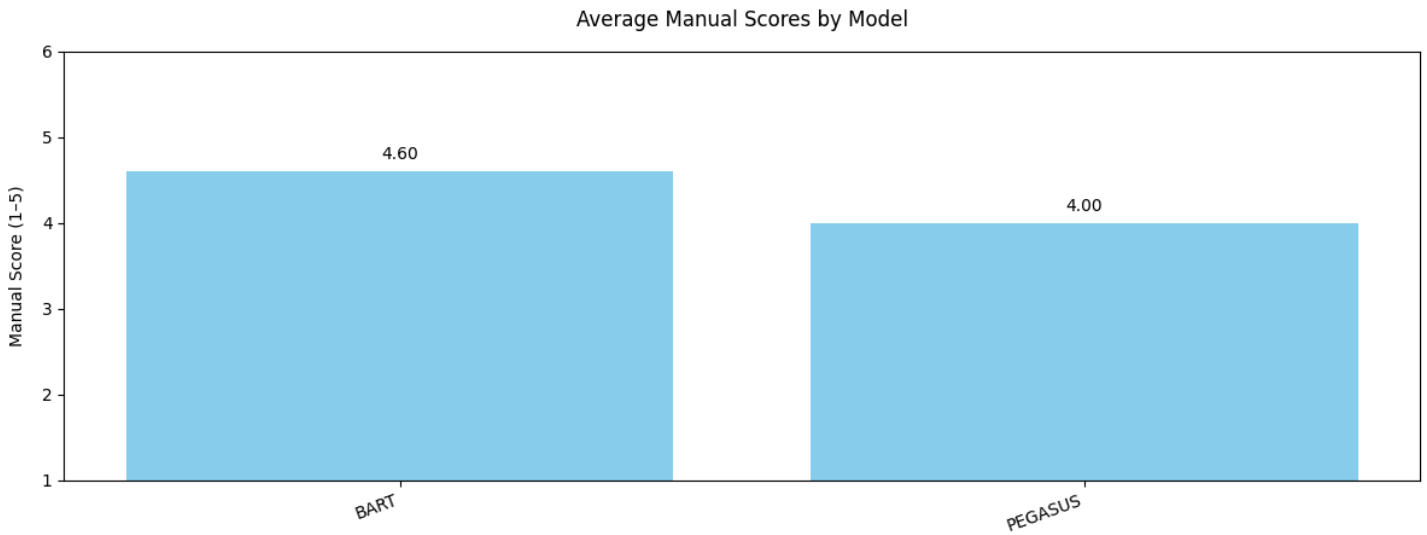
III. RESULTS

A. BART vs PEGASUS

By both intrinsic (ROUGE) and extrinsic (manual) evaluation methods, the default BART summarizer slightly outperformed the PEGASUS summarizer on the CNN/Daily Mail dataset.



BART Default outperformed PEGASUS Default slightly across all ROUGE metrics.



BART summaries were slightly more fluent, contained more detail, and avoided repeated phrases.

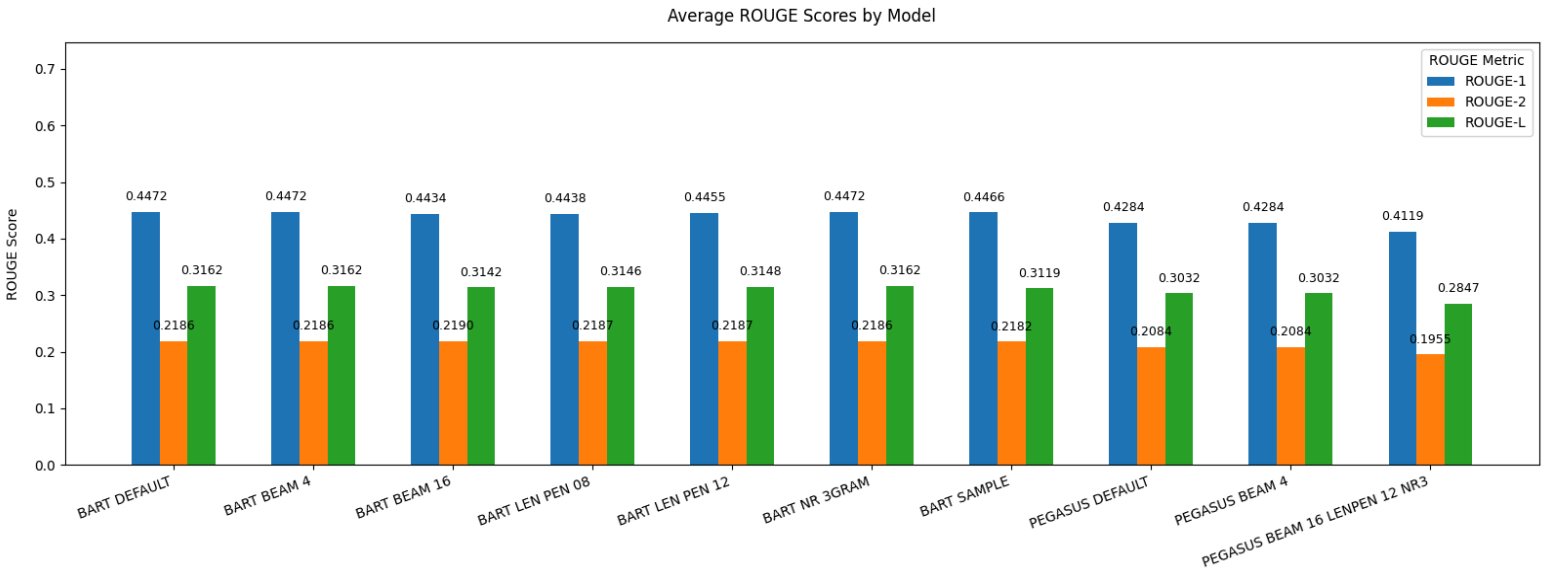
BART's summaries were judged to be slightly more fluent, better at including relevant details, and less prone to repetition. Overall, both models produced summaries of similar quality, with BART showing a slight performance edge. The example summary below highlights mistakes made by PEGASUS that were avoided by BART:

Ground-truth	**BART**	**PEGASUS**
Didier Drogba given the Barclays Spirit of the Game award . *The 37-year-old's foundation has done impressive work in Africa .* *Some of Chelsea's stars attended a charity ball which raised £400,000 .* *CLICK HERE for all the latest Chelsea news .*	*Didier Drogba has been awarded with a Barclays Spirit of the Game trophy. The Chelsea forward set up the 'DidierDrogba Foundation in Africa' He hopes to inspire the next generation of footballers in Africa to fall in love with the game. The 37-year-old scored against Leicester on Wednesday.*	*Didier Drogba has set up the 'Didier Drogba Foundation in Africa'<n>**The 'Didier Drogba Foundation,' contribute financial and material support in education and health including school bags for the school children .<n>The 'Didier Drogba Foundation,' contribute financial and material support in education and health including school bags for the school children .**<n>Chelsea's stars such as Eden Hazard, Petr Cech and Branislav Ivanovic were out in force earlier this month as they raises £400,000 for the foundation at a charity ball .*

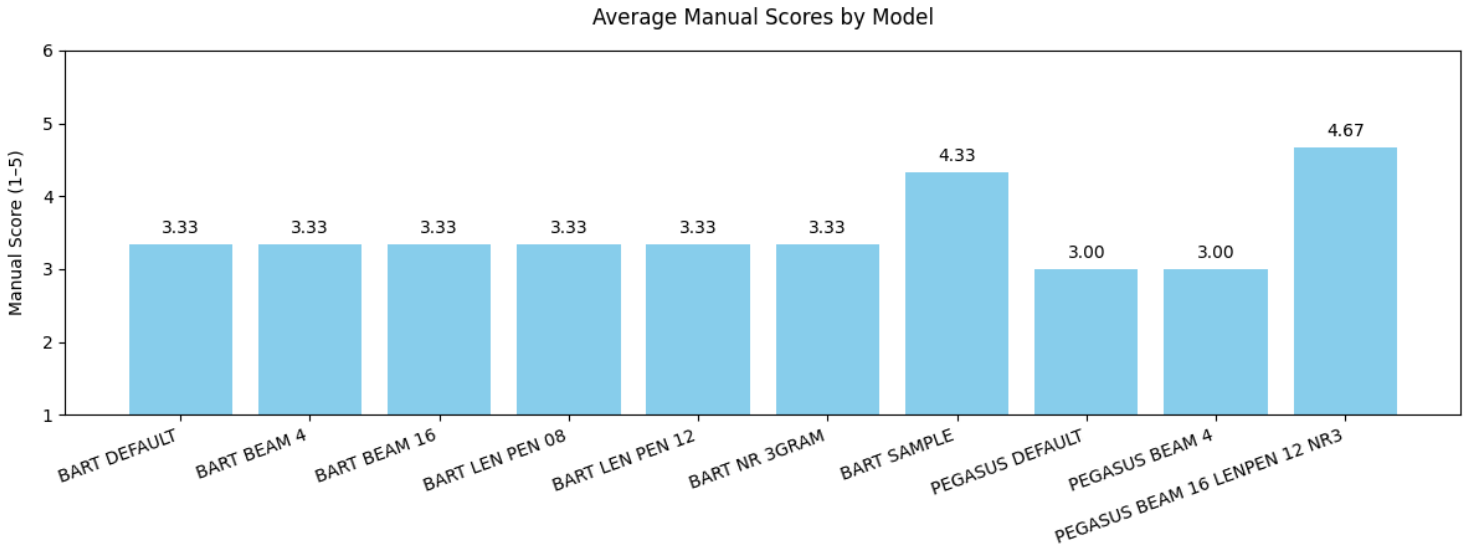
BART produced a higher quality, more succinct summary with more fluent language. PEGASUS captured essential details but generated a clunkier summary with a repeated sentence (red).

B) Ablation Study

Because the default BART model produced slightly higher-quality summaries than the default PEGASUS model in the previous section, the ablation study primarily focuses on exploring hyperparameters to further improve BART's performance.



ROUGE scores remained remarkably stable across all configurations, with no significant improvements observed from varying hyperparameter strategies.



Manual review indicates a slight advantage for the BART sampling approach (top_p = 0.9) and overall superior performance from PEGASUS using beam search (num_beams = 16).

Varying hyperparameter strategies make little difference in overall summary quality, with many outputs between models being identical. The sampling approach, which explores the top 90% of candidates at each decoding step, produces slightly more fluent summaries with more relevant details and resolves some awkward phrasing in both BART and PEGASUS outputs. PEGASUS with aggressive beam search (num_beams = 16) delivered the highest-quality summaries, offsetting some of the shortcomings seen in other configurations. However, the improvement is minor and may only be meaningful to certain users based on personal preference. The example in the table below illustrates the practical differences in summary quality.

Ground-truth	**BART Default**	**BART Sample**	**PEGASUS Beam 16**
Image went viral after it was posted by user Bobitis on Reddit . *The unnamed youth, from New York, knelt on a chair to reach whiteboard .* *The boy is the son of New York professor and drew map from memory .*	*The 11-year-old boy is seen drawing the intricate map of the world **from scratch**. He is the son of a professor in New York who invited him to her class. In the images, the youth is seen kneeling on a chair in order to reach the whiteboard.*	*The 11-year-old is the son of a professor in New York. He is seen kneeling on a chair and drawing the world map **from memory**. **The map is highly accurate, with some major countries labelled. It includes political borders and even tiny coastal islands.***	*The 11-year-old boy **with autism** is seen drawing the intricate map of the world from scratch .<n>In the images, the youth is seen kneeling on a chair in order to reach the whiteboard, with **his pen held to the extremely detailed map**.<n>**Not only is the map accurate, with some major countries labelled, but the high level of detail extends to the inclusion of political borders and even tiny coastal islands .***

The BART Default model captures almost all of the necessary details, but fails to mention that the 11-year-old subject of the article has autism, and that his map is exceptionally detailed. The BART Sample model contains more relevant information (yellow) and better phrasing (green). The PEGASUS Beam 16 mode produced the best summary with all necessary details and better, more fluent phrasing (green).

VI. DISCUSSION

A. Interpretation of results

Both intrinsic and extrinsic evaluations suggest that, while summaries can be slightly improved through sampling and aggressive beam search, in the case of PEGASUS, the resultant summaries differ little in both content and phrasing. These findings are consistent with the original PEGASUS publication, which reported that a robust PEGASUS model only slightly outperforms a comparable BART model on the CNN/Daily Mail dataset. For news summarization, this indicates that model parameters make little difference, and that users may prefer cheaper and quicker configurations over more robust and expensive ones, which offer only modest improvements in output quality.

B. Limitations

The stability of ROUGE metrics across configurations in the ablation study suggests that ROUGE may be less effective as a sole performance metric than desired. Many strengths and weaknesses of BART and PEGASUS configurations only became apparent through manual review, highlighting the need for human evaluation to provide essential context to ROUGE scores.

Key limitations include the computational cost of generating summaries for any configuration and the reliance on manual review for accurate quality assessment. With additional time or computational resources to produce and review more summaries, the configurations explored might have shown greater divergence, potentially leading to more impactful findings.

C. Future Work

This project demonstrates that a robust PEGASUS model can produce slightly better summaries than a comparable BART model. This modest improvement appears to stem from the GSG masking approach, in which the most important sentences from the input text are masked during encoding and predicted during decoding. Outside of this masking strategy, PEGASUS and BART share the same encoder–decoder architecture. This suggests that future work could explore alternative or hybrid masking approaches that might yield more substantial improvements in summary quality. Given the negligible quality differences observed across hyperparameter configurations, further exploration of hyperparameter strategies may not be productive.

V. CONCLUSION

This project provides a head-to-head analysis of BART and PEGASUS abstractive summarization models, both pre-trained on the CNN/Daily Mail dataset. Results show little difference in the quality of summaries produced by the two models and demonstrate that varying hyperparameter configurations has minimal impact on summary quality. Any improvements observed in the ablation configurations were mostly limited to phrasing and structure, with only minor changes to content. These findings suggest that, for real-world applications, the choice between BART and PEGASUS should be guided primarily by personal preference and convenience. Despite the close performance, the GSG masking strategy accounts for some minor improvements over BART summaries when using a robust PEGASUS configuration, indicating that further exploration of new or hybrid masking approaches could yield additional improvements.

REFERENCES

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development, 2*(2), 159–165. https://doi.org/10.1147/rd.22.0159

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop (pp. 74–81). Association for Computational Linguistics. https://aclanthology.org/W04-1013/

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. Advances in Neural Information Processing Systems, 28, 1693–1701. https://arxiv.org/abs/1506.03340

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://arxiv.org/abs/1706.03762

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. https://arxiv.org/abs/1910.13461

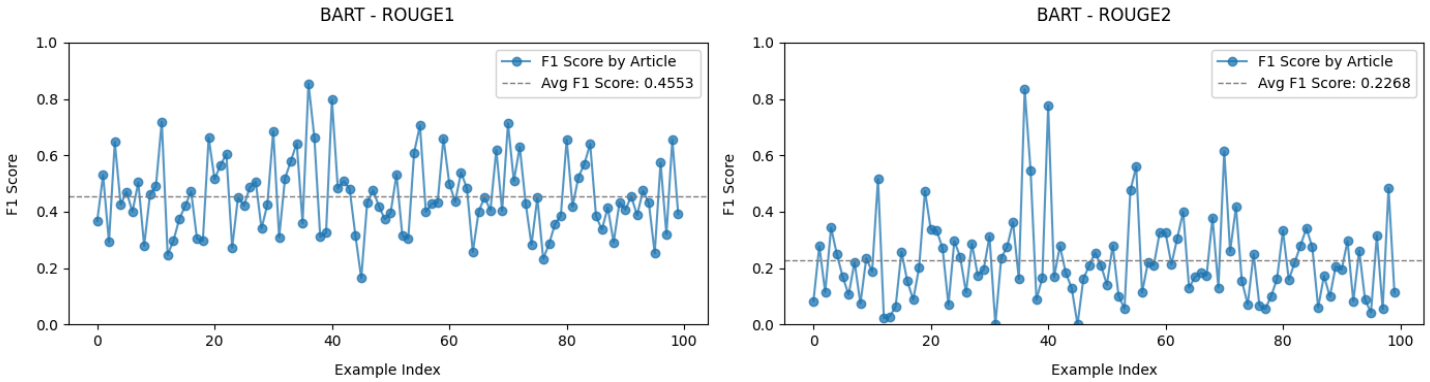
Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2019). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. https://arxiv.org/abs/1912.08777

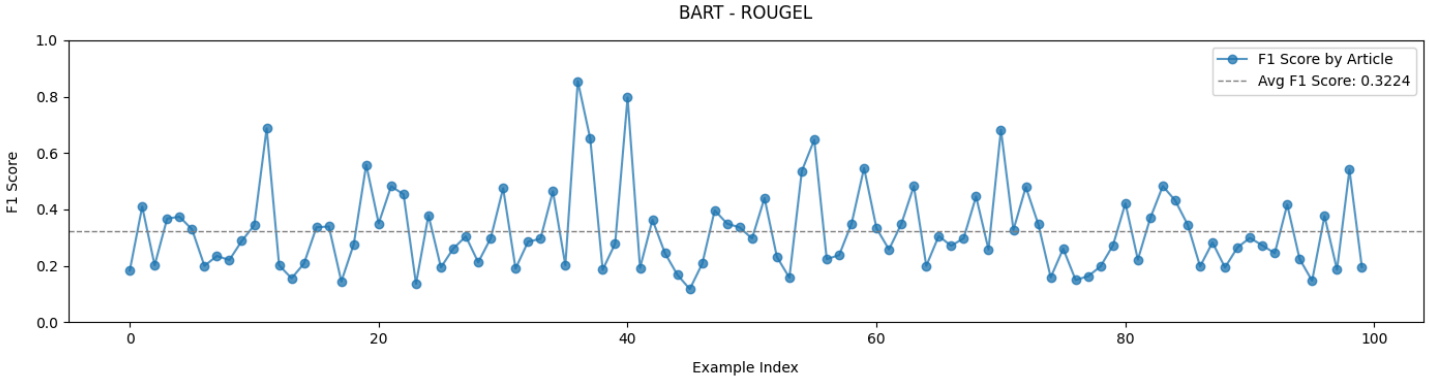
APPENDICES

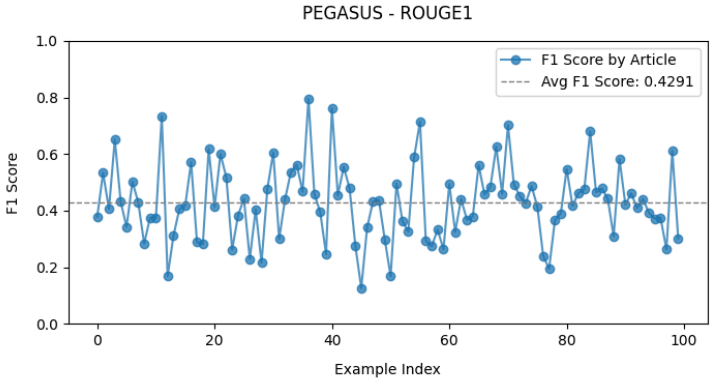
All evaluation data, including model summaries, and code is available on Github

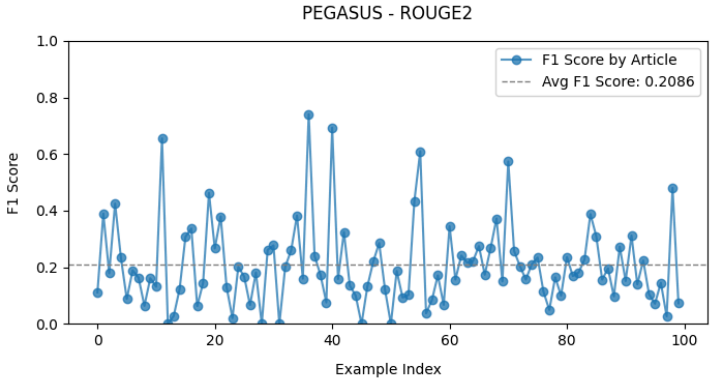
Appendix A - BART vs PEGASUS (Default Configurations)

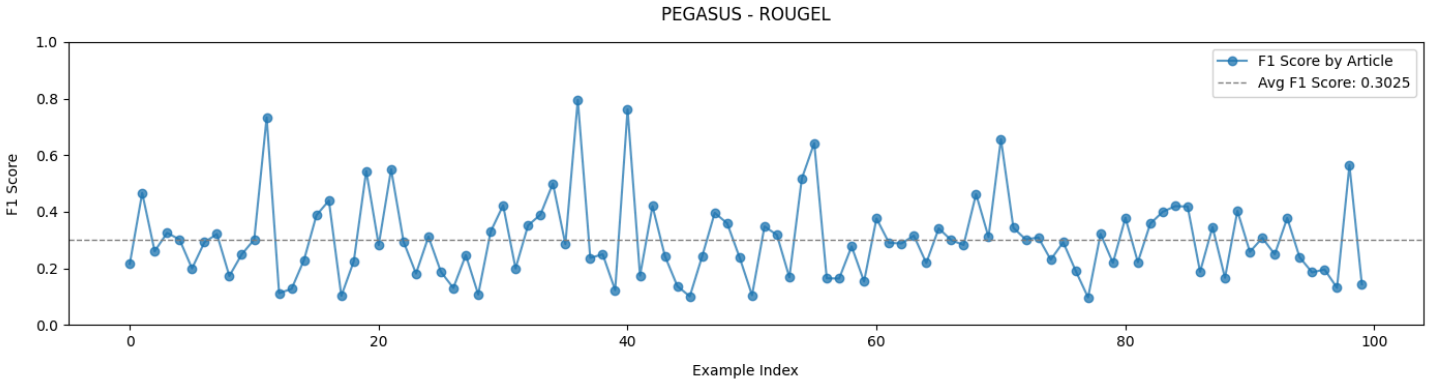
ROUGE-1, ROUGE-2 & ROUGE-L (Averaged Over 100 Summaries)

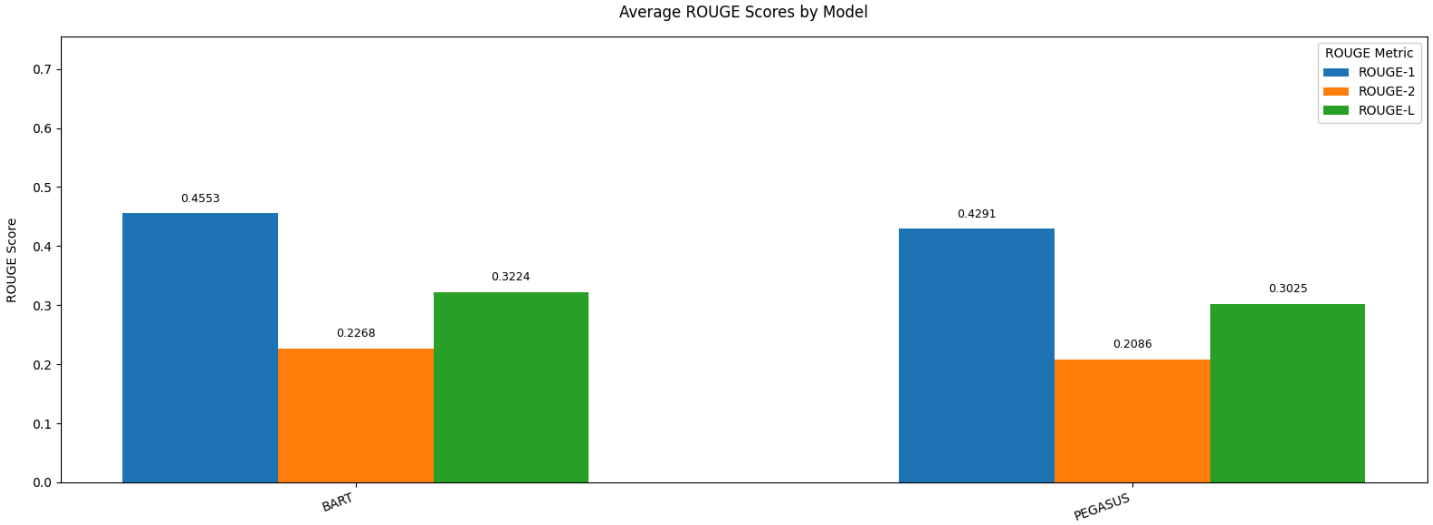






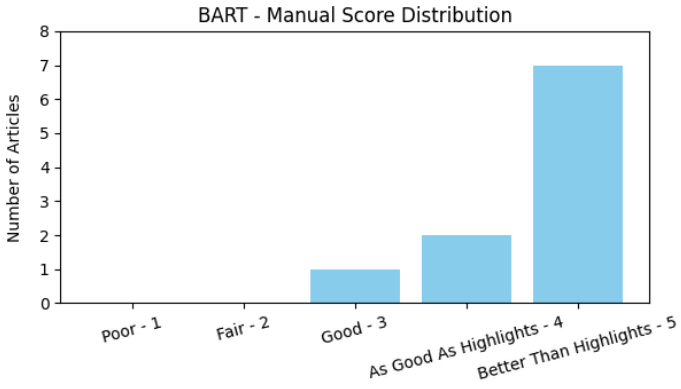


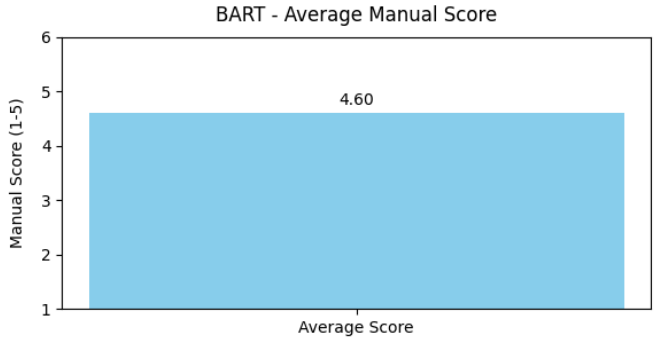


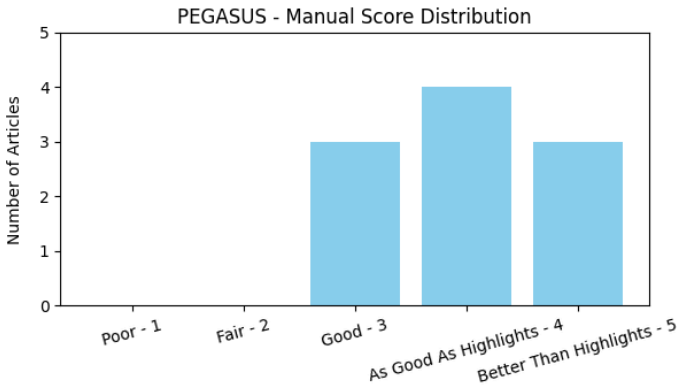


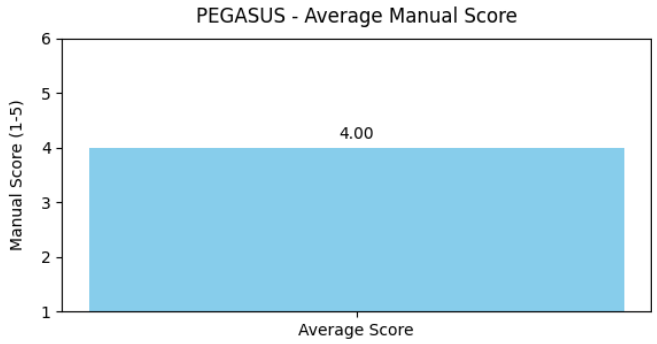
BART slightly outperforms PEGASUS in ROUGE-1 (0.4553 vs 0.4291), ROUGE-2 (0.2268 vs 0.2086) and ROUGE-L (0.3224 vs 3025) metrics.

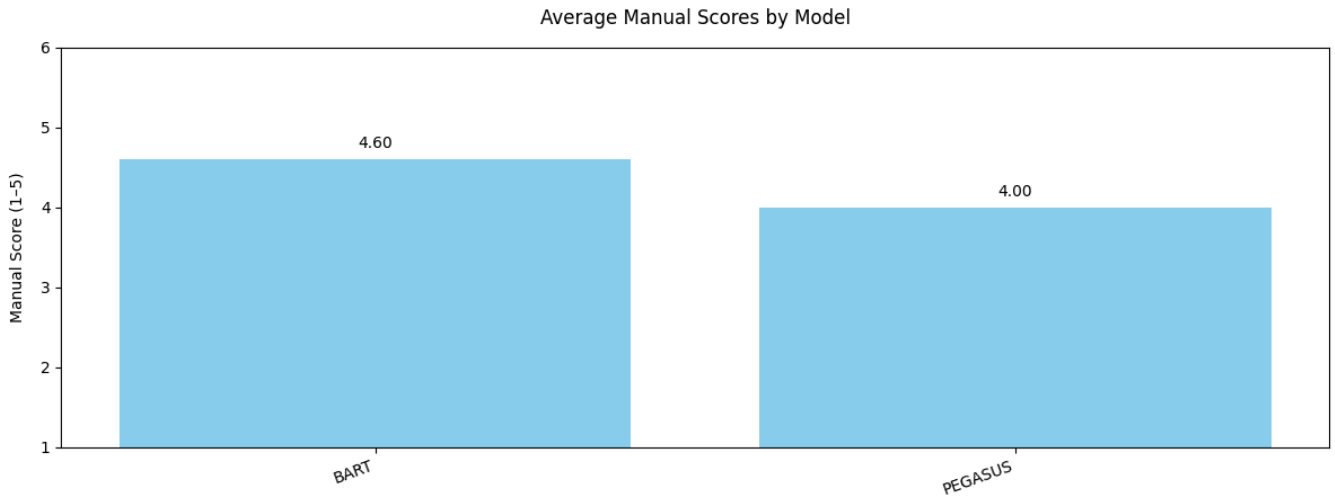
Manual Evaluation (Averaged Over 10 Summaries)











BART outperformed PEGASUS in manual evaluation (average scores: 4.60 vs. 4.00). While both models produced summaries with nearly identical content, BART's summaries were a bit smoother and better structured. PEGASUS occasionally struggled with repeated phrases.

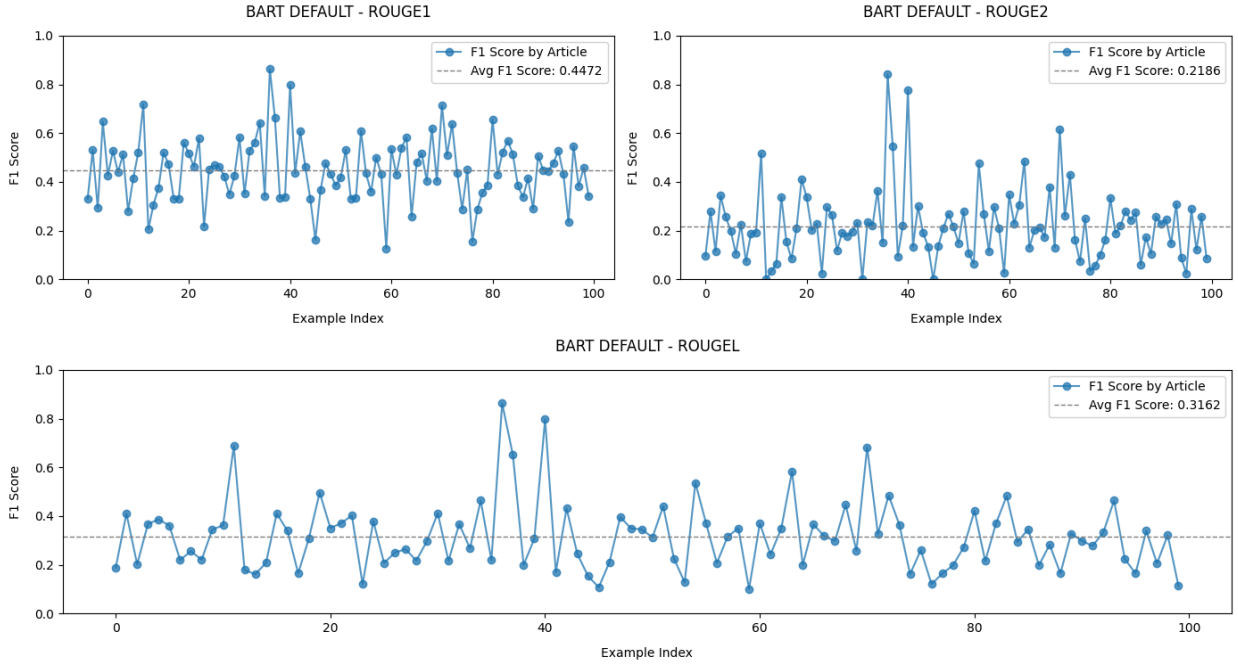
Summary Example

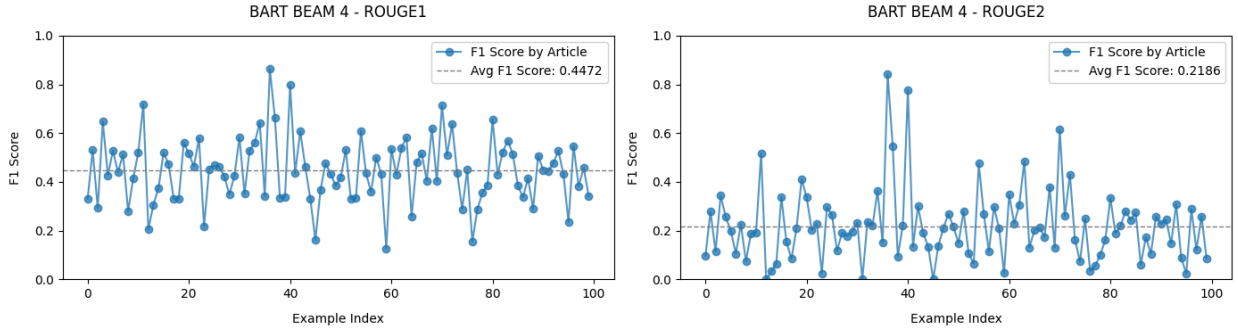
Ground-Truth	**BART (4)**	**PEGASUS (3)**
Cate McGregor is a Group Captain in the Royal Australian Air Force . *She spent 40 years in the army, most of those under the name Malcolm .* *In 2012, McGregor stopped 'functioning' as a man and lived as a woman .* *She tried to resign from the office when her transformation became public .* *Her resignation was refused by former Chief of Army David Morrison .* *McGregor believes Mr Abbott wasn't given the credit he deserved for publicly embracing her and risking a wedge within his own party .* *She addressed the National Press Club on Wednesday as part of a women in media series .*	*Cate McGregor was surprised when she was unconditionally backed by the cricket community and the Australian Army. But it was Tony Abbott's response to her transformation from man to woman that knocked her for six. It took less than an hour for the federal Liberal leader to call Ms McGregor after she forwarded the chapter of her book that outed her. Ms McGregor recalled that very moment - during her address to the National Press Club on Wednesday - when the prime minister told her 'it was okay' 'I actually said to him "your party won't be thrilled about this",' she said.*	*Cate McGregor spoke to the National Press Club on Wednesday .<n>It took less than an hour for the prime minister to call her after she forwarded the chapter of her book that outed her .<n>The senior military official did not express the same kind of appraisal when it came to Mr Abbott's stance on same-sex marriage .<n>Ms McGregor, now Group Captain in the Royal Australian Air Force, spent 40 years in the Army, most of those under the name Malcolm .<n>In 2012, after years of drowning agonising pain with alcohol and drugs, McGregor stopped 'functioning' as a man and chose to live*

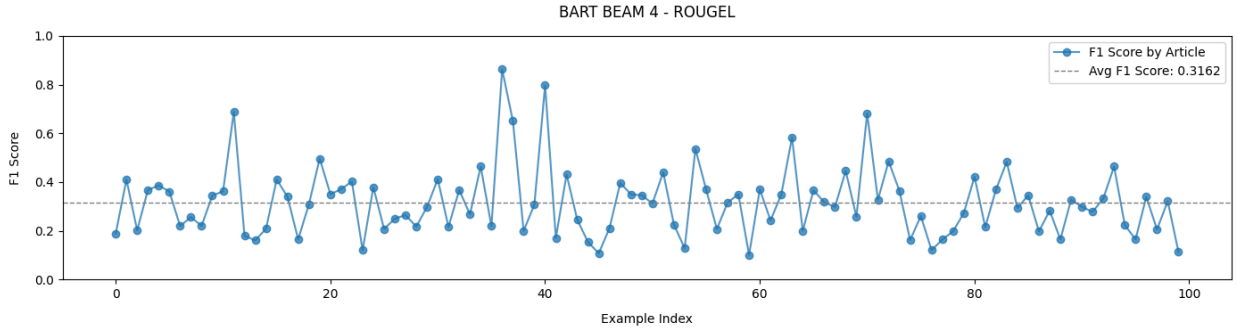
The BART summary is more direct and to the point, while the PEGASUS summary conveys the same information but meanders and leaves the final sentence incomplete. The BART summary received a score of 4, whereas the PEGASUS summary received a score of 3.

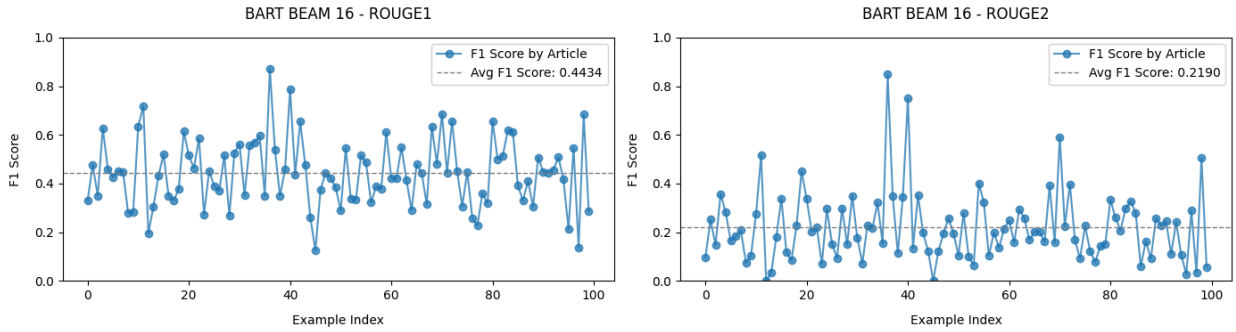
Appendix B - Ablation Study

Ablation Study - ROUGE-1, ROUGE-2 & ROUGE-L (Averaged Over 100 Summaries)

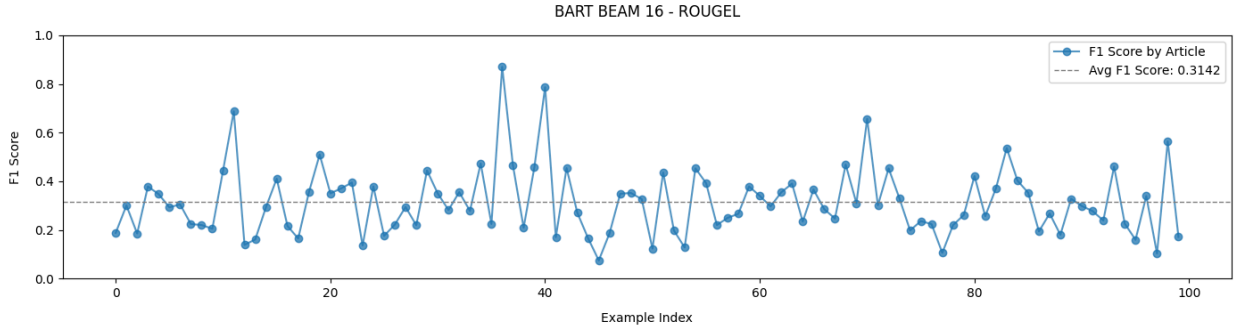


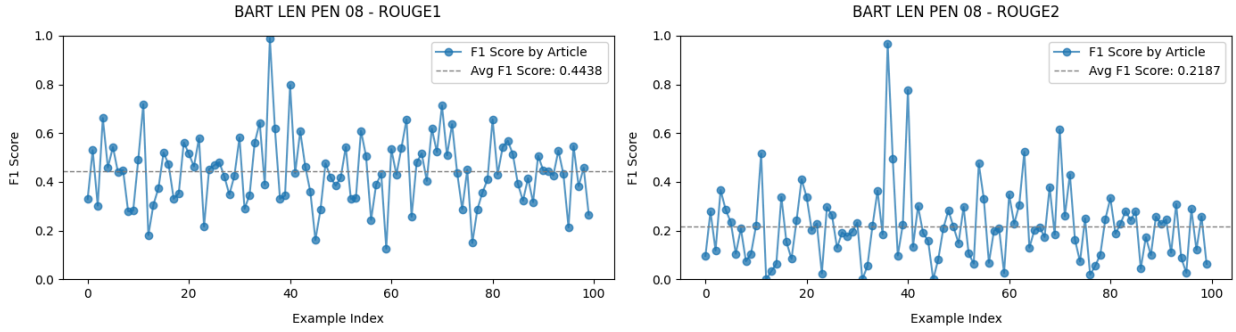




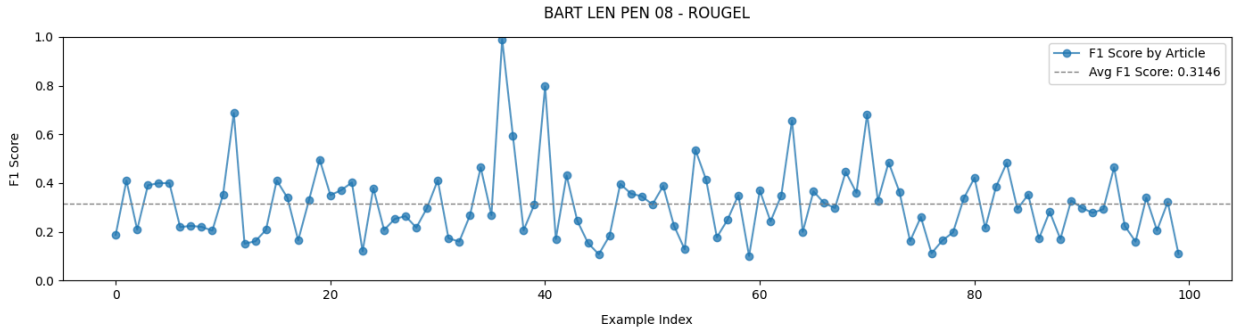

BART BEAM 16 - ROUGE1
F1 Score by Article
Avg F1 Score: 0.4434
F1 Score
Example Index

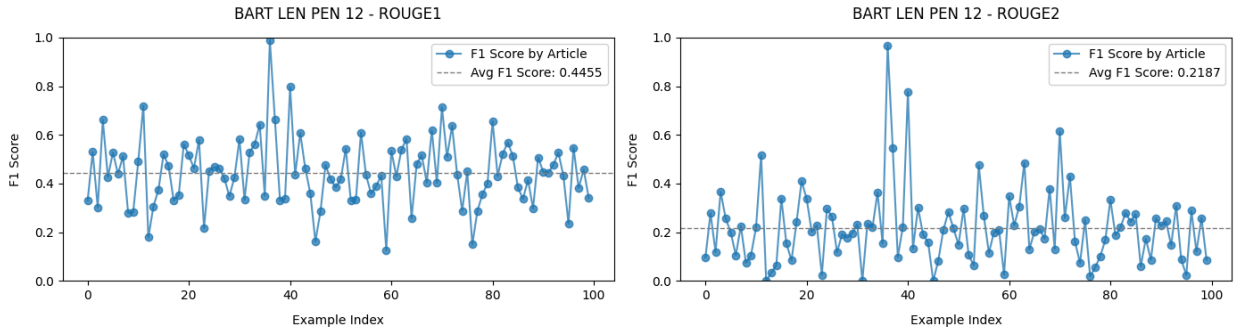
BART BEAM 16 - ROUGE2
F1 Score by Article
Avg F1 Score: 0.2190
F1 Score
Example Index


BART BEAM 16 - ROUGEL
F1 Score by Article
Avg F1 Score: 0.3142
F1 Score
Example Index

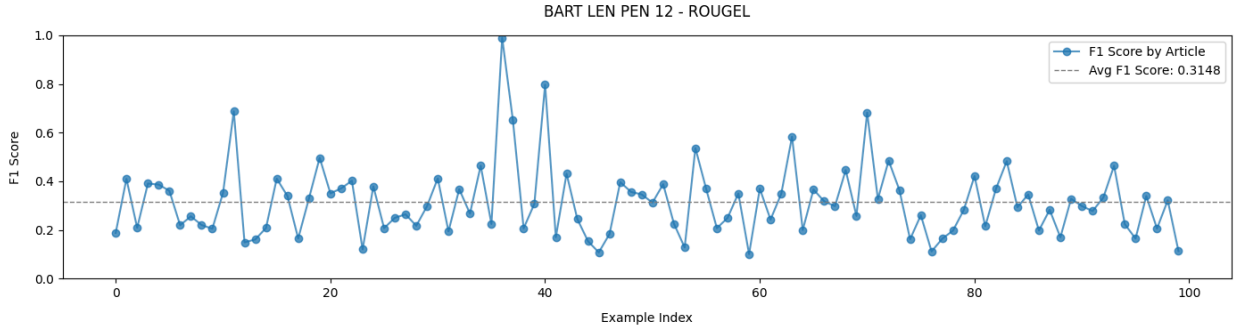

BART LEN PEN 08 - ROUGE1
F1 Score by Article
Avg F1 Score: 0.4438
F1 Score
Example Index

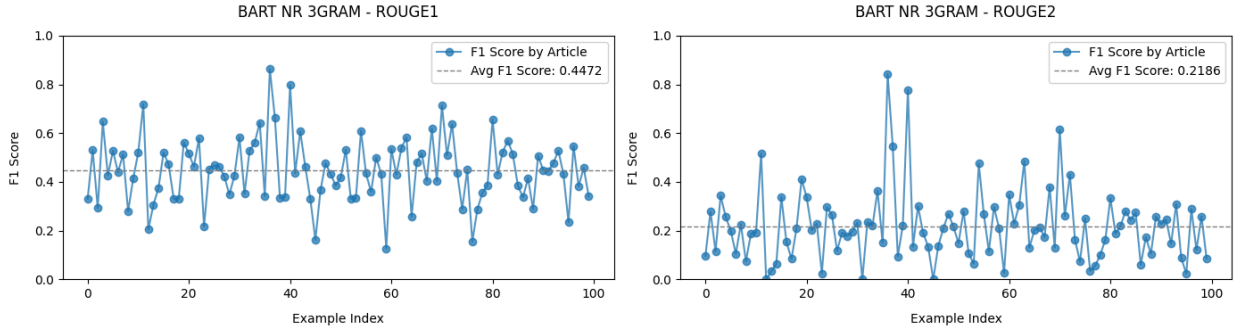
BART LEN PEN 08 - ROUGE2
F1 Score by Article
Avg F1 Score: 0.2187
F1 Score
Example Index


BART LEN PEN 08 - ROUGEL
F1 Score by Article
Avg F1 Score: 0.3146
F1 Score
Example Index

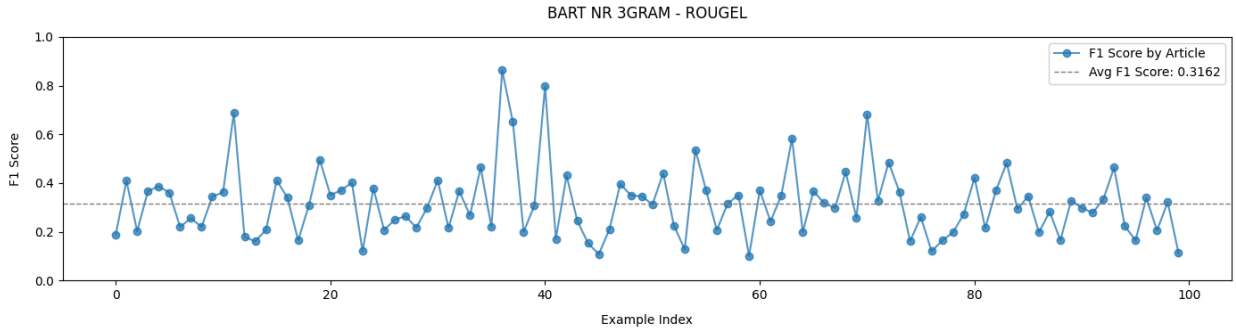

BART LEN PEN 12 - ROUGE1
F1 Score by Article
Avg F1 Score: 0.4455
F1 Score
Example Index

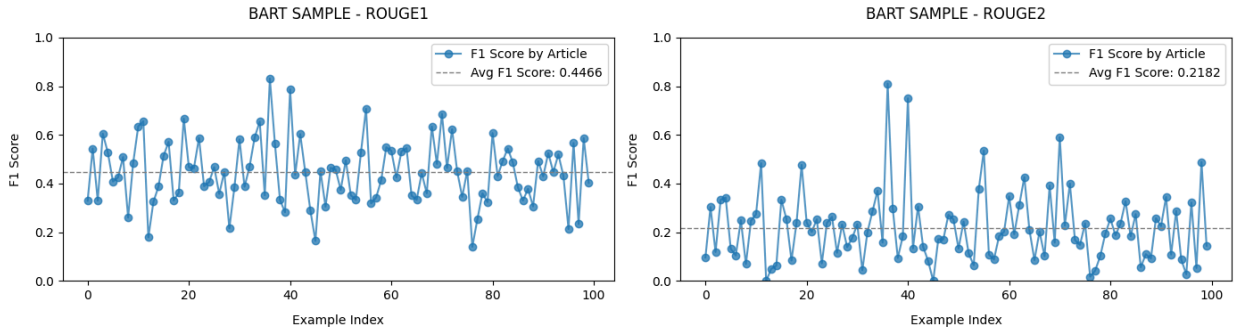
BART LEN PEN 12 - ROUGE2
F1 Score by Article
Avg F1 Score: 0.2187
F1 Score
Example Index


BART LEN PEN 12 - ROUGEL
F1 Score by Article
Avg F1 Score: 0.3148
F1 Score
Example Index

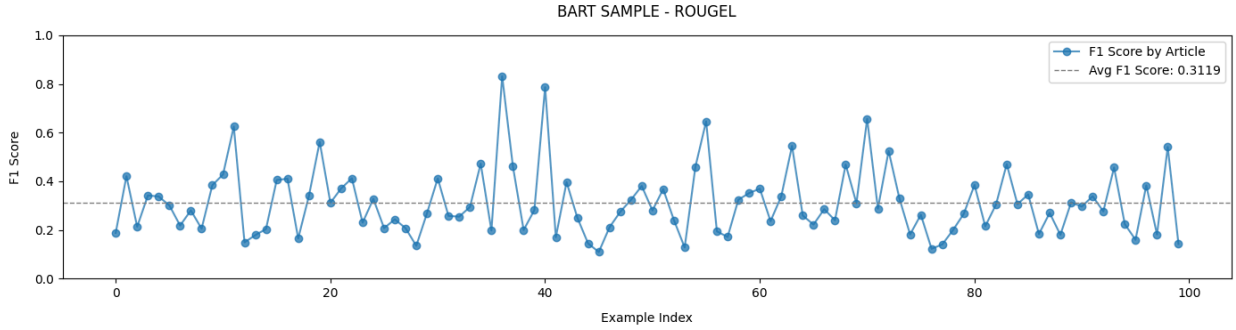

BART NR 3GRAM - ROUGE1
F1 Score by Article
Avg F1 Score: 0.4472
F1 Score
Example Index

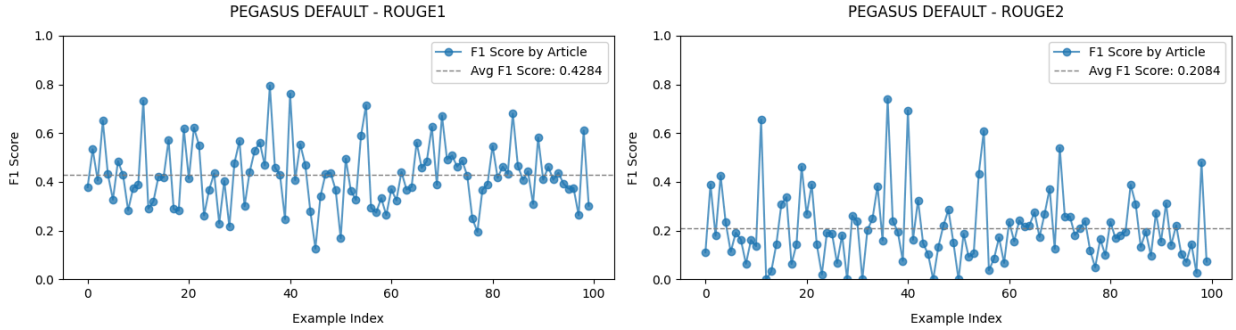
BART NR 3GRAM - ROUGE2
F1 Score by Article
Avg F1 Score: 0.2186
F1 Score
Example Index


BART NR 3GRAM - ROUGEL
F1 Score by Article
Avg F1 Score: 0.3162
F1 Score
Example Index

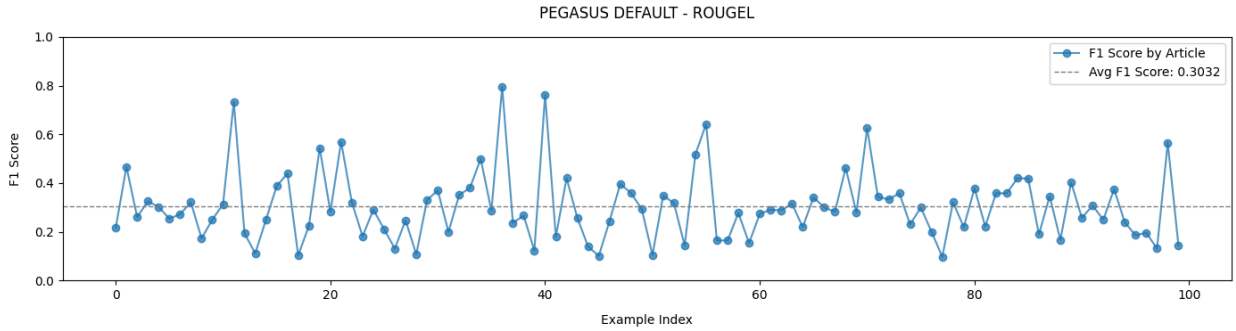

BART SAMPLE - ROUGE1
F1 Score by Article
Avg F1 Score: 0.4466
F1 Score
Example Index

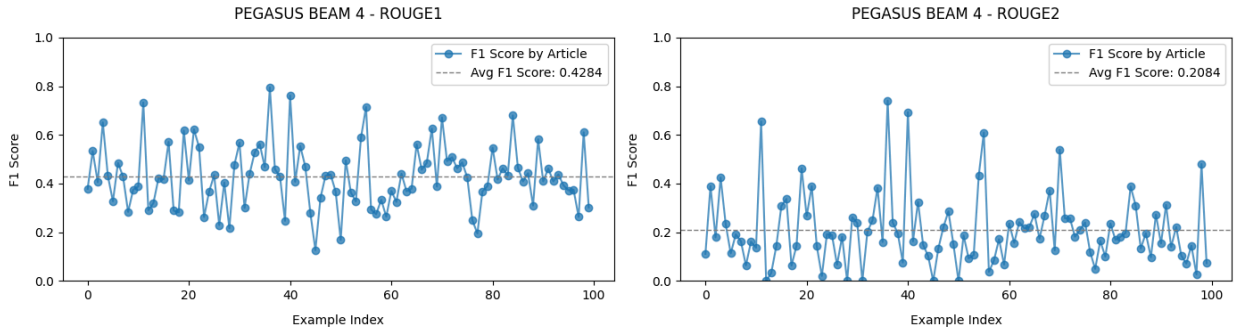
BART SAMPLE - ROUGE2
F1 Score by Article
Avg F1 Score: 0.2182
F1 Score
Example Index


BART SAMPLE - ROUGEL
F1 Score by Article
Avg F1 Score: 0.3119
F1 Score
Example Index

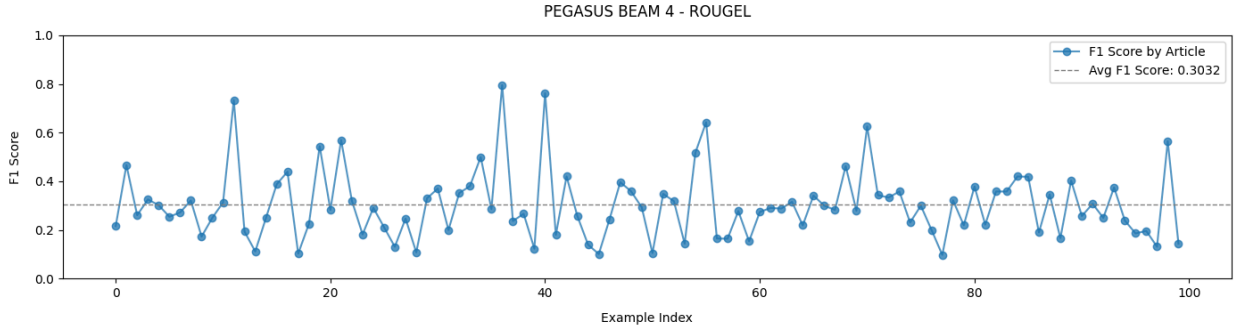

PEGASUS DEFAULT - ROUGE1
F1 Score by Article
Avg F1 Score: 0.4284
F1 Score
Example Index

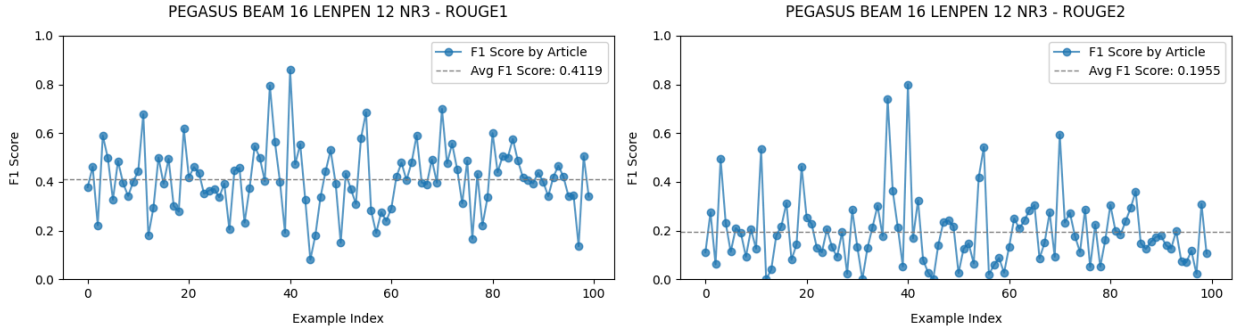
PEGASUS DEFAULT - ROUGE2
F1 Score by Article
Avg F1 Score: 0.2084
F1 Score
Example Index


PEGASUS DEFAULT - ROUGEL
F1 Score by Article
Avg F1 Score: 0.3032
F1 Score
Example Index

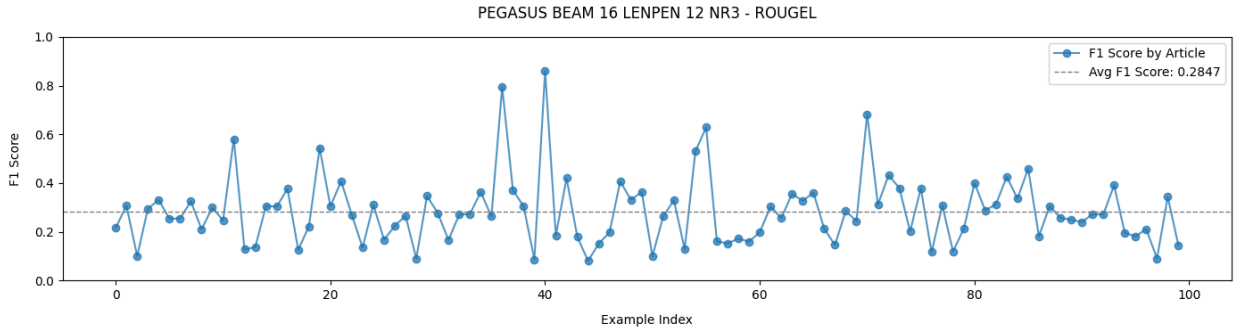

PEGASUS BEAM 4 - ROUGE1
F1 Score by Article
Avg F1 Score: 0.4284
F1 Score
Example Index

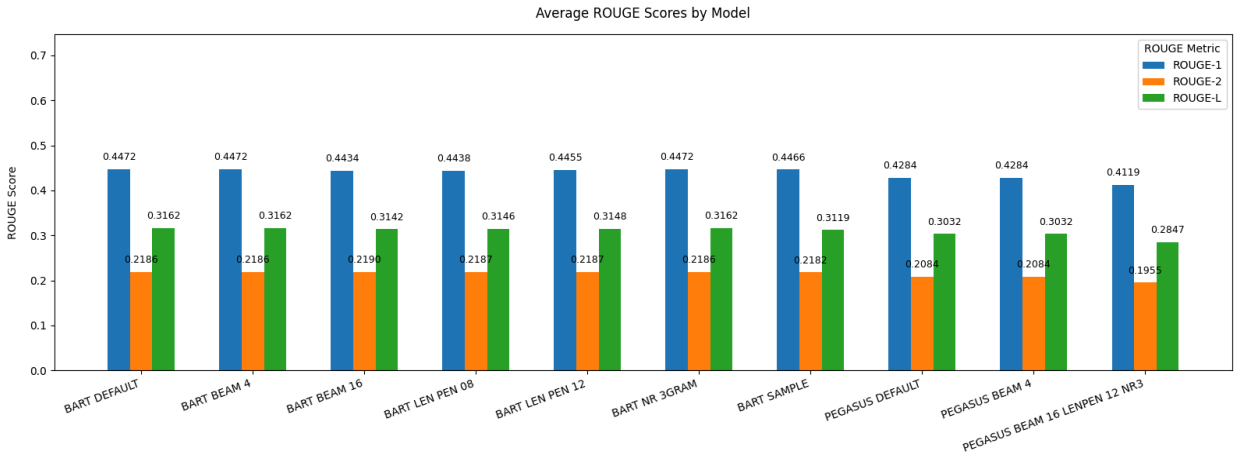
PEGASUS BEAM 4 - ROUGE2
F1 Score by Article
Avg F1 Score: 0.2084
F1 Score
Example Index


PEGASUS BEAM 4 - ROUGEL
F1 Score by Article
Avg F1 Score: 0.3032
F1 Score
Example Index


PEGASUS BEAM 16 LENPEN 12 NR3 - ROUGE1
F1 Score by Article
Avg F1 Score: 0.4119
F1 Score
Example Index

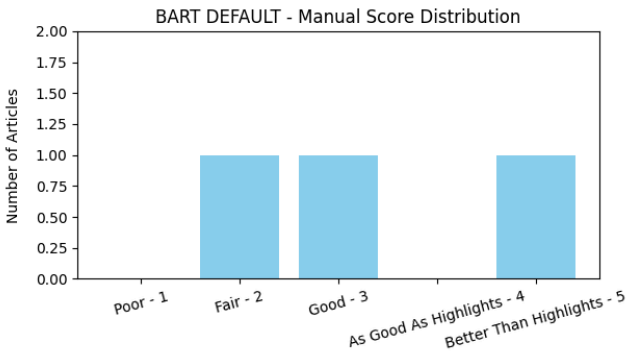
PEGASUS BEAM 16 LENPEN 12 NR3 - ROUGE2
F1 Score by Article
Avg F1 Score: 0.1955
F1 Score
Example Index

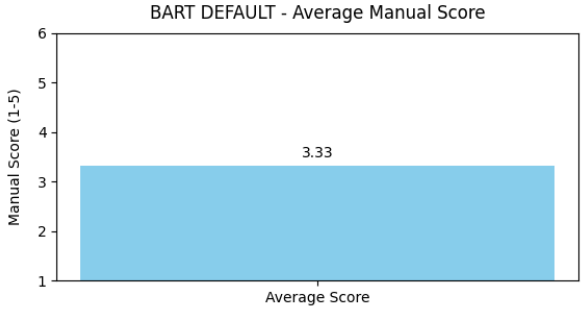

PEGASUS BEAM 16 LENPEN 12 NR3 - ROUGEL
F1 Score by Article
Avg F1 Score: 0.2847
F1 Score
Example Index

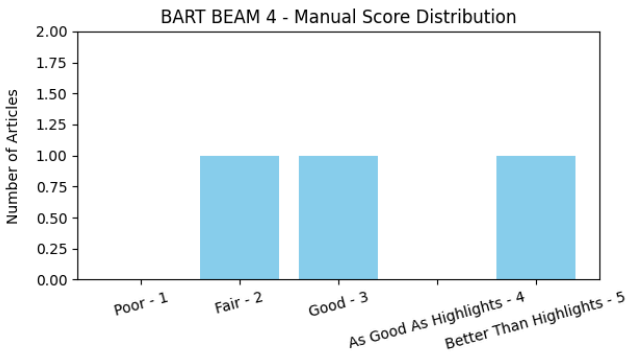


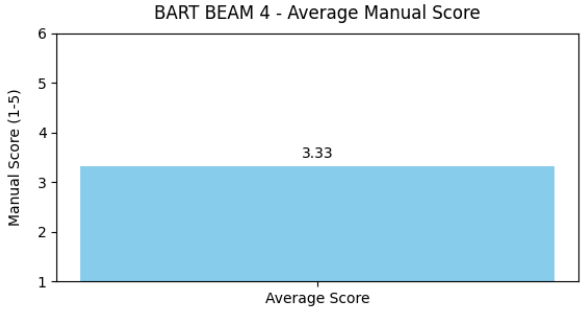
ROUGE scores stayed about the same across all ablation configurations and didn't match the relative differences seen in manual evaluation.

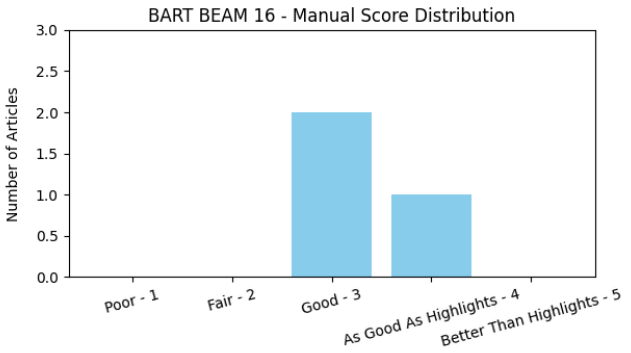
Ablation Study - Manual Evaluation (Averaged Over 3 Summaries)

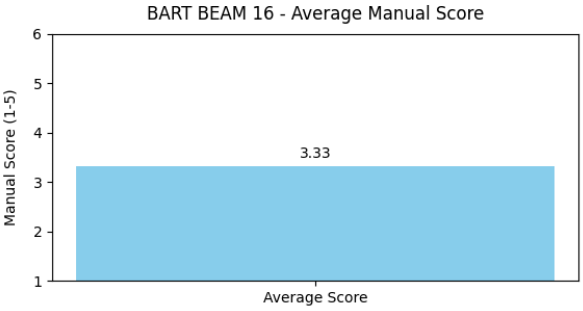


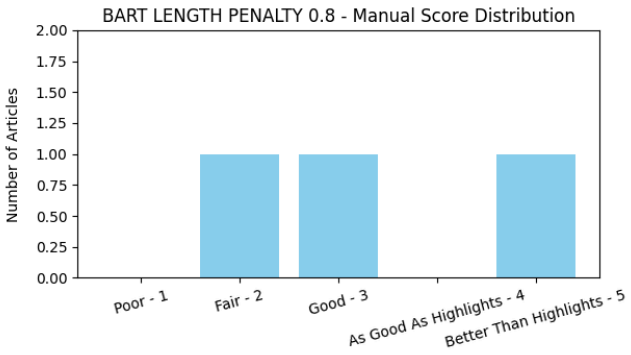


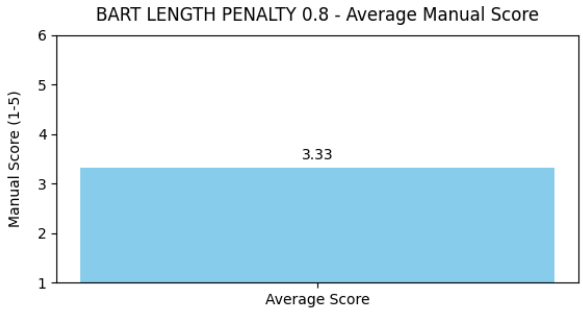


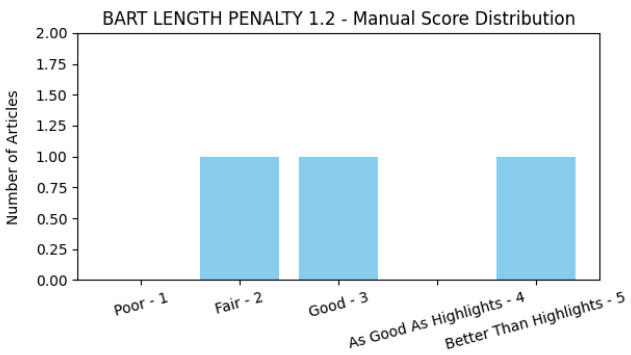


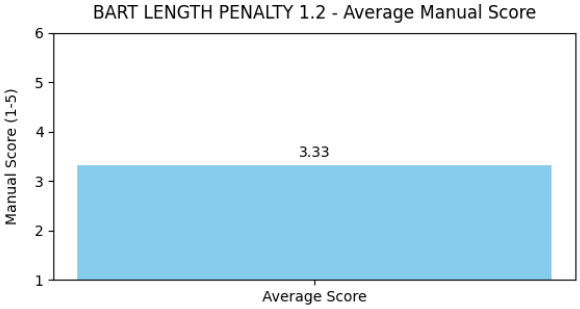


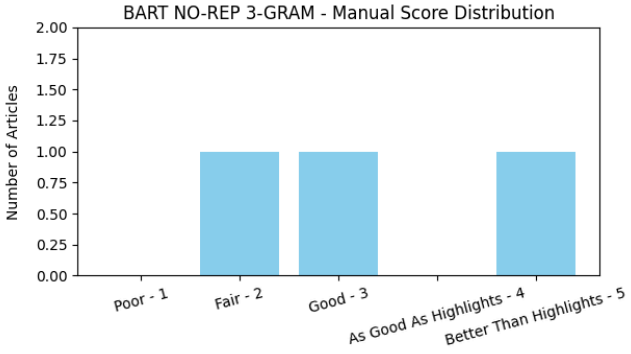


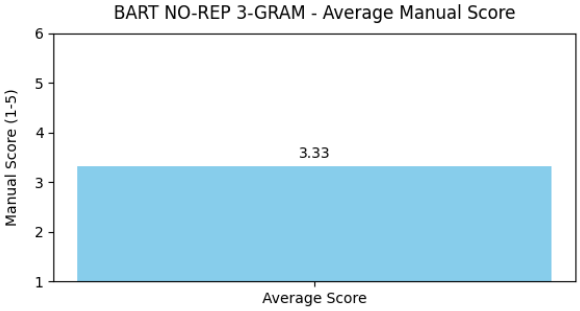


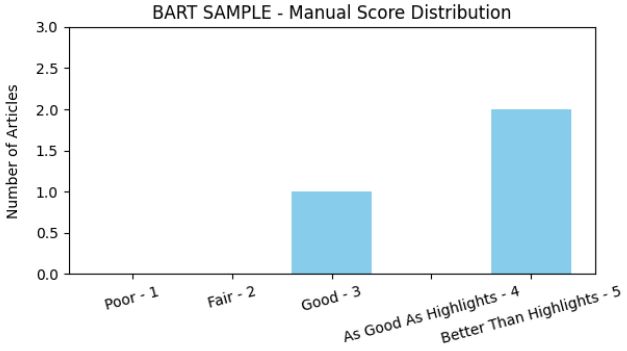


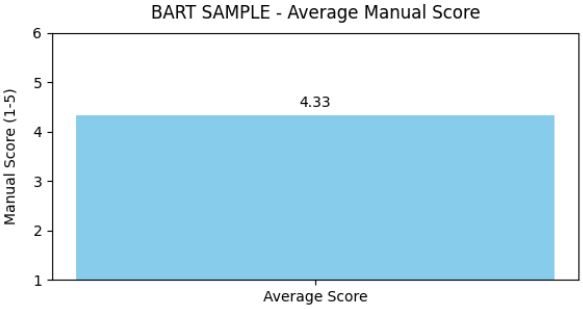

BART LENGTH PENALTY 1.2 - Manual Score Distribution
Number of Articles
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Poor - 1
Fair - 2
Good - 3
As Good As Highlights - 4
Better Than Highlights - 5

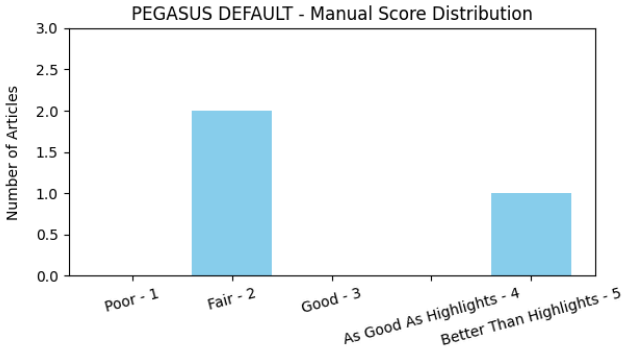

BART LENGTH PENALTY 1.2 - Average Manual Score
Manual Score (1-5)
6
5
4
3
2
1
3.33
Average Score

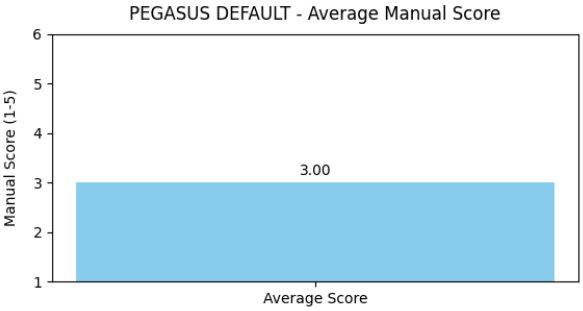

BART NO-REP 3-GRAM - Manual Score Distribution
Number of Articles
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Poor - 1
Fair - 2
Good - 3
As Good As Highlights - 4
Better Than Highlights - 5

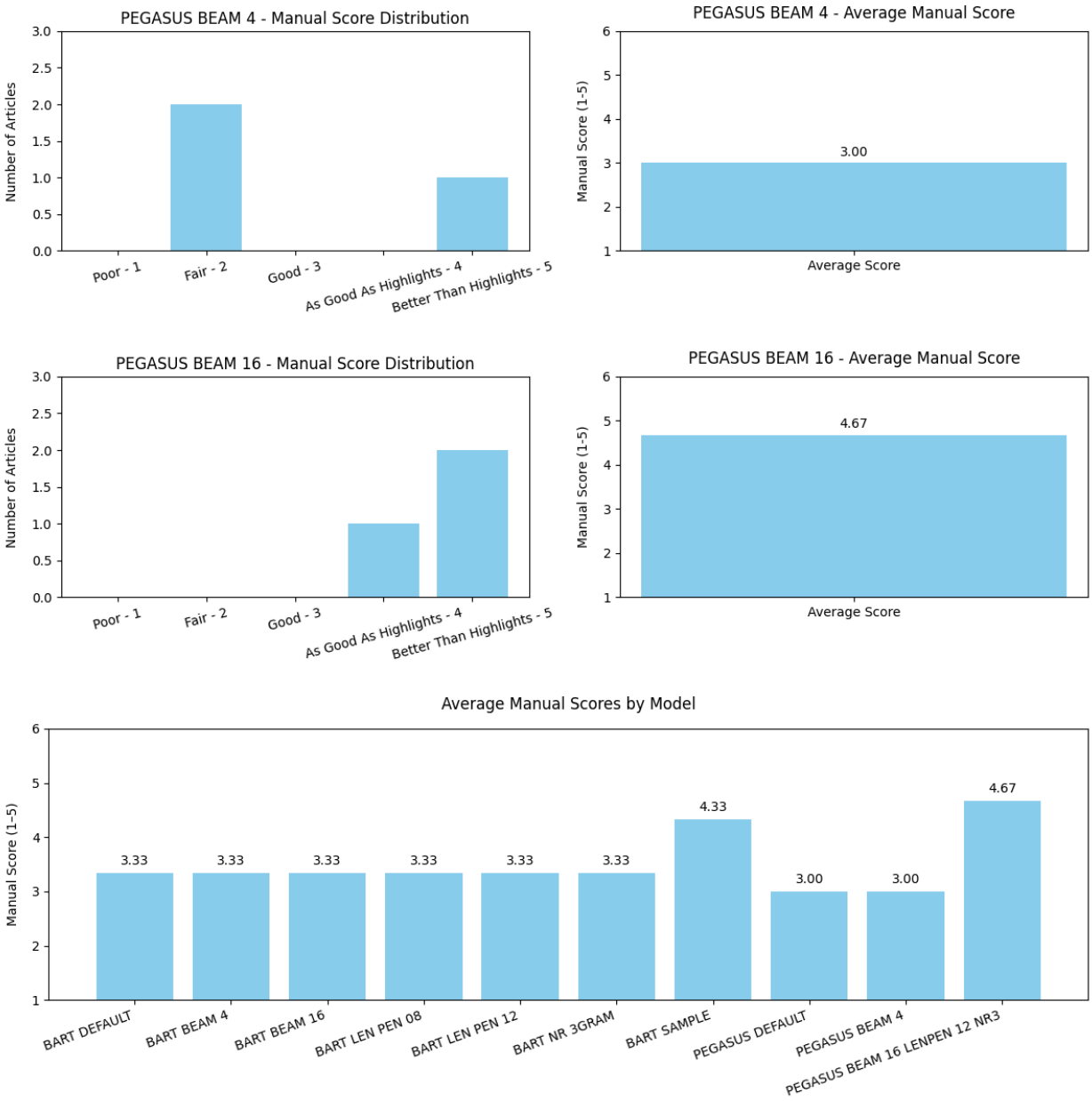

BART NO-REP 3-GRAM - Average Manual Score
Manual Score (1-5)
6
5
4
3
2
1
3.33
Average Score


BART SAMPLE - Manual Score Distribution
Number of Articles
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Poor - 1
Fair - 2
Good - 3
As Good As Highlights - 4
Better Than Highlights - 5


BART SAMPLE - Average Manual Score
Manual Score (1-5)
6
5
4
3
2
1
4.33
Average Score


PEGASUS DEFAULT - Manual Score Distribution
Number of Articles
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Poor - 1
Fair - 2
Good - 3
As Good As Highlights - 4
Better Than Highlights - 5


PEGASUS DEFAULT - Average Manual Score
Manual Score (1-5)
6
5
4
3
2
1
3.00
Average Score



Most BART ablation configs produce nearly identical summaries and scores. Sampling the top 90% of candidates (top_p = 0.9) gives a slight quality boost. PEGASUS with beam search (num_beams = 16) yields the best summaries, but only by a small margin.

Ablation Study - Summary Example

Ground-Truth	BART (Others*) (2)	BART Sample (5)	PEGASUS Beam 16 (4)
Oldham have signed 16-year-old winger Ronaldo Brown . Brown was released by Liverpool and has joined the League One club . He is named after Brazilian Ronaldo and has a brother called Rivaldo . Brown also has a younger sister called Trezeguet .	*Ronaldo Brown has joined Oldham Athletic from Liverpool.* ***The 16-year-old is named after Brazil's former striker Ronaldo. But he is actually named after the original Brazilian Ronaldo, rather than Portugal star Cristiano.***	*Ronaldo Brown has joined Oldham Athletic from Liverpool.* ***The 16-year-old is named after Brazil's former striker Ronaldo. He also has a twin brother called Rivaldo and a younger sister called Trezeguet.***	*Ronaldo Brown has signed for Oldham Athletic after being released by Liverpool .<n>****The 16-year-old is named after Brazil's former striker Ronaldo, who is pictured scoring in the 2002 World Cup.<n>But he is actually named after the original Brazilian Ronaldo, rather than Portugal star Cristiano.****<n>Football-mad mum Denise, 38, chose the name because of her love of the Brazil side - and added that Ronaldo has a twin brother called Rival*

Most BART models generate awkward phrasing for this article, repeating that the subject is named after the Brazilian Ronaldo and not the Portuguese star (yellow). BART Sample fixes this (green), while PEGASUS Beam 16, the top performer overall, makes the same mistake. Most BART models scored 2, BART Sample scored 5, and PEGASUS Beam 16 scored 4.

**Others include BART Default, BART Beam 4, BART Length Penalty 0.8, BART Length Penalty 1.2, and BART No-Rep 3-Gram. BART Beam 16 produces a slightly different summary but still has the same awkward phrasing (yellow).*