DataSci 266 Final Project

Leveraging LLaMA-3 and GPT-mini for High-Fidelity Summarization and Semantic Evaluation of News Articles

Aditya Kumar
adityakumar225@berkeley.edu

Irina Lee
irilee@berkeley.edu

Matthew Shull
shullm@berkeley.edu

Abstract

This project explores the application of a pre-trained LLaMA model to the task of abstractive summarization using a dataset of CNN and Daily Mail news articles. By leveraging the "meta-llama/Meta-Llama-3.1-8B-Instruct" model, we generate concise summaries for a subset of articles and evaluate their quality using two complementary techniques. First, ROUGE metrics assess the overlap between generated summaries and reference texts, providing a quantitative measure of lexical similarity. Second, a modern semantic textual similarity (STS) strategy is implemented, using GPT-4o-mini to rate the semantic alignment between summaries and references on a Likert scale from 0 to 5. The results demonstrate the effectiveness of pre-trained models in producing coherent summaries without additional training while highlighting areas where alignment with human reference summaries can be improved. This study underscores the practicality of combining automated and human-like evaluation metrics for comprehensive performance assessment in natural language processing tasks. Results showed that without using a zero-shot LLaMA model, generated summaries were robust but differed from human structure.

1. Introduction

1.1. Problem Statement: Automatic text summarization has become a vital tool in an era of information overload, offering a way to distill large volumes of text into concise and meaningful summaries. High-quality summaries are essential for enhancing comprehension, supporting decision-making, and improving accessibility to textual content across various domains. Despite advancements in natural language processing, generating abstractive summaries that are both semantically accurate and linguistically coherent remains a significant challenge. Most state-of-the-art summarization models require custom training on domain-specific datasets, which demands extensive computational resources, time, and expertise. This creates a barrier for practical applications, especially when working with new or diverse datasets.

1.2. Project Motivation: Pre-trained models, such as the LLaMA family, offer a promising alternative by eliminating the need for additional hyperparameter-tuning. This project evaluates the applicability of a LLaMA model for summarizing news articles from the CNN and Daily Mail News dataset, focusing on its ability to generate coherent and meaningful summaries without domain-specific fine-tuning. By applying modern evaluation techniques, including ROUGE metrics and semantic textual similarity (STS) scoring via GPT-4o-mini, this study investigates the practical capabilities of Large Language Models on unseen datasets. The insights gained from this evaluation could help bridge the gap between research-grade NLP models and real-world summarization tasks, offering a scalable solution for applications in media and beyond.

1.3. Hypothesis: Leveraging pre-trained LLaMA based models for text summarization will result in summaries that demonstrate high semantic coherence, lexical overlap (as measured by ROUGE scores), and alignment with human-generated reference texts.

2. Literature Review

In the past, the use of pre-trained models like PEGASUS and LLaMA has shown significant promise in addressing the challenges of abstractive summarization. The paper by Zhang et al. (2023) highlights the performance of fine-tuned models on various datasets, demonstrating their ability to capture semantic meaning with minimal distortion when paired with task-specific training. Their work underscores the importance of scaling and pre-training on diverse datasets for generalizability and performance, aligning with the goals of this project, which evaluates a pre-trained model on unseen data without additional fine-tuning. Additionally, “The Current State of Summarization” by Fabian Retkowski provides an in-depth analysis of the shift toward leveraging pre-trained instruction-tuned models, such as FLAN-T5 and InstructGPT, for summarization tasks. These models excel in zero-shot and few-shot scenarios, reducing the need for fine-tuning while maintaining competitive performance. This aligns with

our approach of using a LLaMA model for summarizing CNN/Daily News articles without additional task-specific training.

In regards to evaluation, evaluating abstractive summarization requires a multidimensional approach to assess both lexical overlap and semantic fidelity. Liu et al. (2023) introduce G-Eval, a framework utilizing GPT-4o-mini to evaluate natural language generation tasks, including summarization, through human-aligned scoring. Their work demonstrates that GPT-4-based evaluation correlates strongly with human judgments, providing an advanced and scalable method for assessing semantic textual similarity (STS). This study directly informed our methodology by guiding the design of STS evaluations using GPT-mini to complement traditional metrics like ROUGE. Fabian Retkowski's paper, "The Current State of Summarization," emphasizes the limitations of traditional metrics like ROUGE in capturing semantic nuances, highlighting the growing importance of modern evaluation techniques such as GPT-based STS scoring. These insights reinforced our decision to integrate GPT-mini evaluations alongside ROUGE for a more comprehensive assessment of summary quality.

The literature consistently highlights limitations in both models and evaluation techniques. Fine-tuned models often face difficulties generalizing to diverse datasets, and metrics like ROUGE fail to capture deeper semantic alignment. As noted in "The Current State of Summarization," pre-trained instruction-tuned models address some of these challenges by eliminating the need for task-specific training. However, their performance can still vary significantly depending on prompt design and input quality. Similarly, Liu et al. (2023) caution that GPT-4 evaluations, while effective, may introduce biases or inconsistencies, particularly when applied to diverse or complex textual domains.

This project builds upon these studies by using a LLaMA model for abstractive summarization and combining ROUGE and GPT-mini STS evaluations to provide a holistic assessment of performance.

3. Dataset

The dataset used for this project consists of CNN and DailyMail news articles, paired with human-generated abstractive summaries (287113 rows by 3 columns). Each article provides a detailed narrative of current events, while the corresponding summaries encapsulate the core information in a concise format. These reference summaries serve as the ground truth for evaluating the performance of the pre-trained LLaMA model. The dataset spans various topics, ensuring diversity in content and complexity, which poses a meaningful challenge for abstractive summarization.

3.1. Preprocessing

To prepare the dataset for model input and evaluation, we implemented the following data cleaning steps:

- Text Cleaning: We developed a clean_text function to standardize the text and remove potential noise. This function removed non-ASCII characters and replaced excessive whitespace with single spaces.

The cleaning process ensured consistency while retaining key punctuation and capitalization, critical for maintaining grammatical accuracy and coherence in the input and reference summaries.

3.2. Sampling Strategy:

- Given the computational demands of the LLaMA model, we randomly sampled 2000 articles from the dataset. This approach provided a sufficient and balanced subset, enabling meaningful evaluation while staying within our computational limits and token limits.

4. Methodology

4.1. Model Selection and Setup:

The "meta-llama/Meta-Llama-3.1-8B-Instruct" model was chosen for its capability to generate high-quality responses without requiring additional fine-tuning. This LLM is specifically designed for instruction-following tasks, making it suitable for abstractive summarization. The 8B parameter variant

was selected to balance computational efficiency and performance, ensuring that the model could handle the complexity of summarizing diverse news articles within our available resources.

4.2. Summarization Process:

The summarization process was designed to generate concise, high-quality summaries from the news articles in the dataset. The workflow included three major stages: input preparation, generation configuration, and post-processing.

4.3. Input Preparation:

The input articles were preprocessed to standardize their structure and ensure compatibility with the model. This step involved cleaning the text by removing non-ASCII characters, excessive whitespace, and noise while preserving essential punctuation and capitalization to maintain grammatical accuracy. Each cleaned article was paired with a well-crafted prompt that guided the model to focus on summarization. The prompts followed a structured format:

- A system message defining the model's role as a summarization expert.
- A user message providing the article text and explicitly requesting a concise summary within four to five lines.

This structure ensured consistency in the inputs and provided clear instructions to the model, improving the quality of the generated summaries.

4.4. Generation Configuration:

The summarization process utilized the LLaMA model. Key parameters were adjusted to optimize the output quality:

- Maximum Token Length: Limited the number of tokens in the output to ensure brevity and relevance.
- Temperature: Controlled the diversity of the generated text by balancing randomness and coherence.
- Top-p Sampling: Applied nucleus sampling to filter out unlikely completions and focus on high-probability responses.

These configurations allowed the model to generate coherent summaries while avoiding overly verbose or irrelevant outputs.

4.5. Post-Processing:

After generating the summaries, a post-processing step was implemented to ensure the outputs were clean and ready for evaluation. This included:

- Stripping any unnecessary artifacts, such as special tokens or incomplete sentences.
- Formatting the summaries to align with the expected grammatical and structural standards.
- Flagging any outputs that failed to meet the expected quality for further review or refinement.
- The finalized summaries were then stored for evaluation, along with their corresponding reference summaries.

5. Evaluation

To assess the quality of the generated summaries, two complementary evaluation metrics were used: ROUGE and Semantic Textual Similarity (STS).

5.1. ROUGE Metrics:

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) was employed to measure the lexical overlap between the generated summaries and the reference summaries. Three specific variants of ROUGE were calculated:

- ROUGE-1: Unigram overlap for content alignment.
- ROUGE-2: Bigram overlap to assess coherence and fluency.
- ROUGE-L: Longest common subsequences for structural similarity.

ROUGE is a widely adopted metric in summarization research due to its ability to quantify content preservation and fluency. Aggregated ROUGE scores across all samples provided a macro-level view of the model’s performance.

5.2. Semantic Textual Similarity (STS)

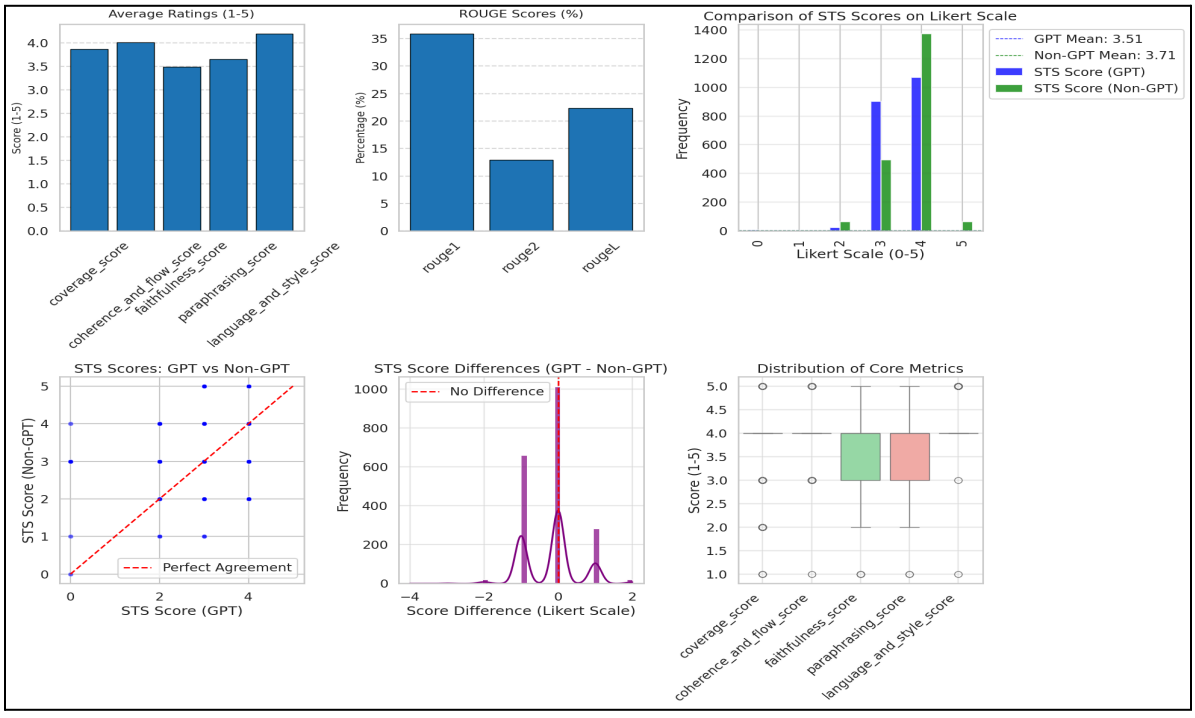
STS evaluated semantic alignment between generated and reference summaries. Using GPT-4o-mini, semantic similarity was scored on a 0-5 Likert scale, focusing on meaning rather than exact wording. Summaries were scored on five weighted criteria: Coverage of Key Ideas (30%), Coherence and Logical Flow (20%), Faithfulness to the Reference (30%), Paraphrasing and Rewording (15%), Language and Style (5%), GPT-4o-mini provided detailed feedback for each category, integrating weighted scores into a final semantic score out of 100.

ROUGE quantified content preservation, while our prompted GPT STS scores captured broader semantic alignment. This combined framework ensured a comprehensive assessment of the summaries, balancing quantitative metrics with qualitative semantic insights. Summaries were analyzed to identify strengths, trends, and areas for improvement in model performance.

Additionally, we used the all-MiniLM-L6-v2 model, a widely used scoring technique to calculate cosine similarity scores between sentence embeddings and mapped these scores to a Likert scale of 0-5 for comparison with a GPT-based scores. This allowed us to directly compare the semantic similarity assessments of both models on a standardized scale

6. Results and Discussion

Scores (to nearest 0.01)	Mean	Standard Deviation	Median	Minimum	Maximum
Coverage of Key Ideas	3.87	0.44	4	1	5
Coherence and Logical Flow	4.00	0.23	4	1	5
Faithfulness to the Reference	3.50	0.62	4	1	5
Paraphrasing and Rewording	3.65	0.51	4	1	5
Language and Style	4.12	0.40	4	1	5
GPT STS Score (0 - 5)	3.51	0.56	4	0	4
Benchmark STS Score (0 - 5)	3.71	0.59	4	0	5
Rouge1 (0 - 1)	0.36	0.09	0.3	0.04	0.6
Rouge2 (0 - 1)	0.13	0.07	0.1	0	0.4
RougeL (0 - 1)	0.22	0.06	0.2	0.04	0.5



The overall results of the study show a strong similarity between the abstract summaries from LLaMA to the human generated summaries in the given dataset. The column chart in the top left shows the average score of the categories created, and the box-and-whiskers plot in the bottle right shows the distribution of the scores and their outliers. Three of the key categories that measure a summarization (coverage of key ideas, coherence and flow, and language and style) are centralized around a measure of 4 with very little dispersion to other scores. The LLaMA model that was used was also reinforced with human responses through instruct-tuning, otherwise known as reinforcement learning with human feedback (RLHF). This is likely why the average score of the 'language and style' category of the LLaMA summary was high. However, the 'faithfulness to the reference' (3.4) and 'paraphrasing and rewording'(3.6) scores indicate that the model may face difficulties in accurately representing the original content while paraphrasing effectively. Inaccuracies in preserving the essence of the original text, as well as challenges in rewording without changing meaning, highlight areas where the summarization model may need further tuning or improved training to the human summaries. There is also the possibility that the human summaries do not fully capture the article white the LLaMA articles capture more of the article.

The ROUGE scores also serve as an indication of the summarization power of the LLaMA model, comparing n-grams to the human generated summary. With the average ROUGE-1 score being the highest, 0.36, the summarization of LLaMA proved to have captured the key unigrams of the human score. The lower scores of ROUGE-2 and ROUGE-L may be attributed to the prompts given to human summarization differing from the summarization prompt given to LLaMA. The low ROUGE-2 score (0.13) suggests that the model struggles with maintaining meaningful bigram (word pair) overlaps, which commonly impacts the readability and natural flow of the summaries. Similarly, the ROUGE-L score (0.22) indicates that while the model does capture some sequential structure, it struggles to consistently preserve the order of terms and phrases in a way that reflects the reference summaries' organization. This is consistent with the expectations of the study since the LLaMA prompt was not fine-tuned on the human-generated summaries and purely used zero-shot. Additionally, the people who created these summaries likely have differing backgrounds in summarization, affecting the ROUGE-2 and ROUGE-L scores heavily when comparing against the LLaMA summary.

The GPT-based model scores slightly lower than the STS Benchmark, indicating that the GPT model may be a stricter grader. This suggests that while the generated summaries are relatively close to the reference summaries in terms of similarity, the GPT model applies a more conservative or stringent evaluation, potentially penalizing subtle differences or discrepancies that may not be flagged by the STS Benchmark. However, the fact that the two mean scores are relatively close implies that the summarization model is performing in line with the standard benchmark, suggesting that the model is producing summaries with comparable quality to human-generated reference summaries. The values of the STS Benchmark from the Sentence Transformers package are created based on a cosine similarity between the human-summaries and the LLaMA-summaries, which focuses more on the terms used in each of the summaries rather than the sentence structure. This is consistent with the results from the ROUGE scores where the ROUGE-1 score is high (0.36) and the LLaMA model has many of the same important unigrams as the human summaries, but lacks similar sentence structure.

7. Conclusion

Overall, the results indicate that the generated summaries maintain strong coherence, style, and coverage, while faithfulness is moderately high but still a point of possible improvement. ROUGE scores suggest moderate lexical overlap, and the STS comparisons reveal that while LLaMA summaries achieve solid semantic alignment, there is a slight difference, but overall our methodology is consistent. Taken together, the metrics and visualizations demonstrate generally favorable performance, with room for refinement in ensuring consistent factual accuracy and alignment with reference texts.

References

Fabian Retkowski. (2023). *The current state of summarization*. Retrieved from https://arxiv.org/pdf/2406.06608

Hou, S., Huang, X., Fei, C., et al. (2021). A survey of text summarization approaches based on deep learning. *Journal of Computer Science and Technology, 36*(3), 633-663. https://doi.org/10.1007/s11390-020-0207-x

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries.

Liu, H., Gao, J., Zhang, X., & Sun, M. (2023). G-Eval: A GPT-based framework for evaluating text summarization models. *arXiv preprint*. Retrieved from https://arxiv.org/abs/2303.12796

Nallapati, R., Zhou, B., Gulcehre, C., et al. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. *arXiv preprint*. Retrieved from https://arxiv.org/abs/1602.06023

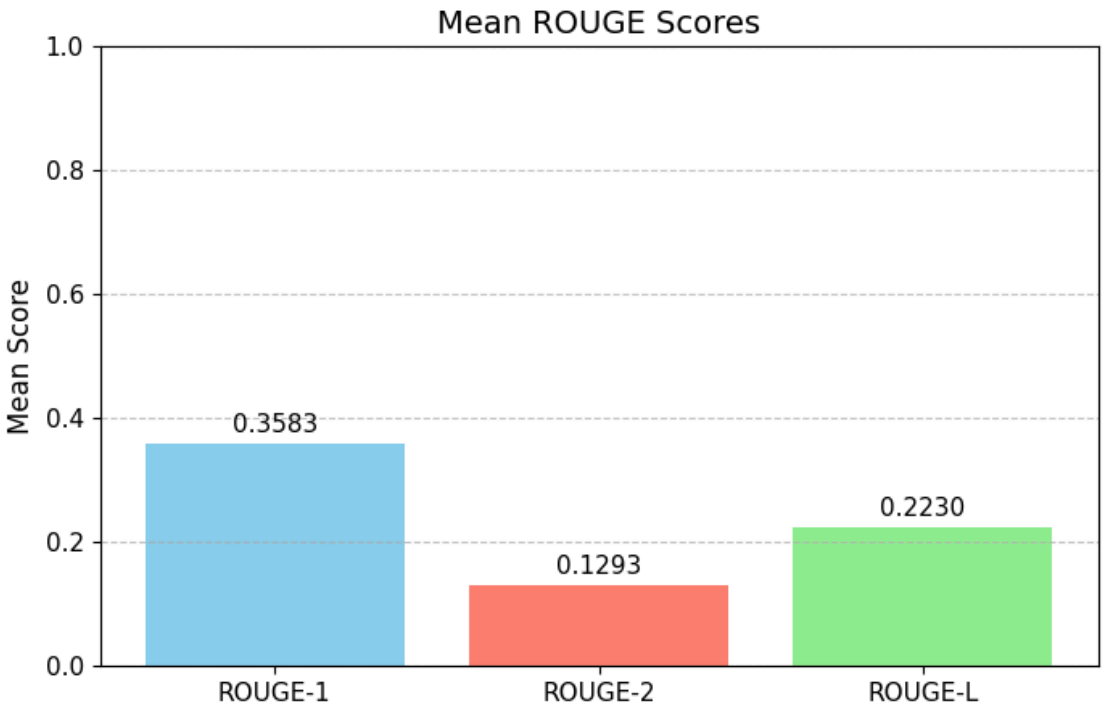
Paulus, R., Xiong, C., & Socher, R. (2017). A deep reinforced model for abstractive summarization. *arXiv preprint*. Retrieved from https://arxiv.org/abs/1705.04304

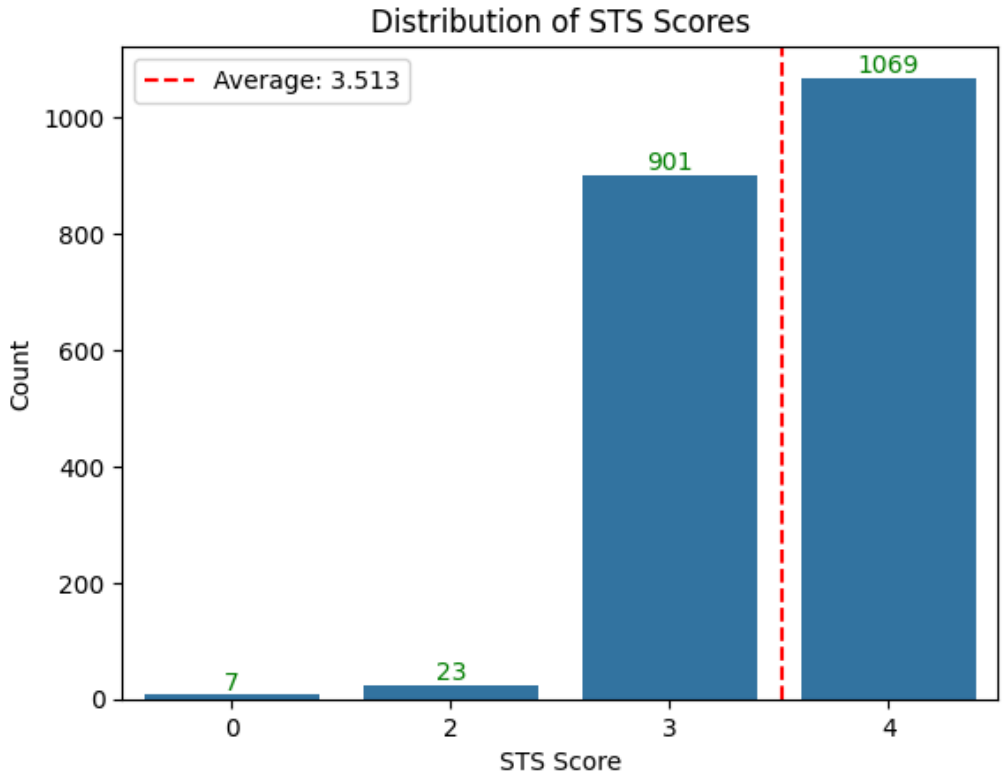
Rehman, T., Ahmed, Z., & Khan, F. (2023). Evaluating pre-trained models for abstractive text summarization: Challenges and opportunities. *arXiv preprint*. Retrieved from https://arxiv.org/abs/2303.16634

Yadav, D., Gautam, S., Sharma, K., et al. (2022). Feature-based automatic text summarization methods: A comprehensive state-of-the-art survey. *IEEE Access*. https://doi.org/10.1109/access.2022.3231016

Zhang, Z., Liu, P., & Wang, Y. (2023). A review of language model scaling for abstractive summarization. *arXiv preprint*. Retrieved from https://arxiv.org/abs/2406.06608

APPENDIX


Mean ROUGE Scores
Mean Score
1.0
0.8
0.6
0.4
0.2
0.0
0.3583
0.1293
0.2230
ROUGE-1
ROUGE-2
ROUGE-L


Distribution of STS Scores
Average: 3.513
Count
1000
800
600
400
200
0
7
23
901
1069
0
2
3
4
STS Score

Sample of Feedback output:

'feedback': {'coverage': 'The generated summary covers most key points but misses the detail about him being a father-of-two.', 'coherence_and_flow': 'The summary flows well and presents information in a logical order.', 'faithfulness': 'The summary accurately reflects the events but adds some details not present in the reference.', 'paraphrasing': 'The paraphrasing is effective, maintaining the original meaning while using different wording.', 'language_and_style': 'The language is clear and professional, with no grammatical errors.'}}