Multi-Document Summarization on Medical Studies

Maryam Feizabad **Kemalcan Jimmerson**
University of California, Berkeley
maryam.feizabad@berkeley.edu kemalcan@berkeley.edu

ABSTRACT
Reviews of medical studies serve as very valuable tools for researchers but are expensive to produce manually. With this project, we aimed to develop a model that creates automated summaries of multiple medical studies. We fine-tuned multiple large language models and compared the results. We found the best results were generated using the LongformerEncoderDecoder (LED) model.

INTRODUCTION
This project was designed to address the growing challenges in medical research caused by the increasing volume of medical literature. It aimed to automate the summarization of medical reviews across multiple documents, as medical studies often contain vast amounts of data with conflicting or inconclusive findings. The manual summarization of such data and studies are time-consuming and resource-intensive. By employing advancements in natural language processing (NLP), the project sought to streamline the extraction of key insights from extensive medical literature, thus alleviating the burden on researchers and healthcare professionals and enabling quicker distillation of crucial information.

The integration of automated summarization and utilizing large language models could significantly enhance the analysis and interpretation of medical text by generating concise, coherent summaries and tackle issues like redundancy and contradictions in the data. Which eventually aimed to enhance efficiency and accuracy in research and healthcare decision-making. To achieve this objective, multiple language models were employed. In the initial approach we fine tuned baseline models to establish performance benchmarks. Subsequent efforts were directed towards refining the summarization process by exploring various models and techniques to improve the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores—a widely used metric for evaluating summarization quality. Notably, among the models evaluated, the large LED model emerged as the most promising candidate, exhibiting superior performance in terms of ROUGE score and overall summarization quality.

The Multi Document Summarization for Literature Review (MSLR) dataset from Hugging Face was utilized as the primary dataset for training and evaluating the summarization models, providing a diverse and comprehensive collection of medical literature for analysis.

BACKGROUND
While significant research exists in the realm of Multi-Document Summarization (MDS), particularly in open-domain settings, there is a notable gap regarding the summarization of medical reviews across multiple documents.
One notable study, "Open Domain Multi-Document Summarization: A Comprehensive Study of Model Brittleness under Retrieval", delves into the challenges of open-domain MDS. This research underscores

the necessity for innovative retrieval and summarization methods, highlighting the importance of factors such as the number of retrieved documents for effective summarization. However, its focus primarily lies outside the domain of medical literature.

In our effort to automate medical review summarization, we draw upon foundational papers such as "BART: Denoising Sequence-to-Sequence Pretraining for Natural Language Generation, Translation, and Comprehension" and "BioBART: Pretraining and Evaluation of a Biomedical Generative Language Model". These seminal works introduce key concepts in sequence-to-sequence pretraining, serving as cornerstones in the development of advanced language models. Additionally, insights from the LoRA (Low Rank Adaptation for Large Language Models) paper inform our approach, offering methodologies to enhance model interpretability and performance.
While existing literature on MDS provides valuable insights, the unique nature of medical review summarization necessitates tailored solutions. Our study aims to bridge this gap by developing novel approaches informed by established methodologies and techniques. By leveraging advancements in natural language processing and drawing upon the rich foundation of MDS research, we seek to address the specific challenges posed by automating medical review summarization.

METHODS

The MS2 (Multi-Document Summarization of Medical Studies) dataset in the biomedical domain includes input strings averaging 39,818 characters, ranging from 263 to 729,579, and tokens averaging 6,873, with extremes from 43 to 123,039. The authors tackled large inputs by using a custom transformer architecture that segmented the input through an overlapping window technique before encoding. These segments were then concatenated post-encoding to prepare for decoding.

We established our baseline model with T5. Then we fine-tuned several large models like BioBart and LED. We iteratively refined our summarization approach by incorporating advanced techniques such as Low Rank Adaptation for Large Language Models (LoRA), LoRA+, TF-IDF, and experimenting with various hyperparameters and configurations. We trained multiple configurations, initially using the ROUGE metric to select promising models for their ability to capture key source text information. Further refinement was conducted through a hyperparameter search, using BERT-scores to fine-tune the semantic quality of summaries to achieve an optimal minimal summary.

T5

Fine-tuning the T5 model yielded only limited success due to constraints in training with only 1000 tokens, which led to memory capacity issues. The memory requirements escalate as the size of the input increases because the dimensions of the transformer's attention heads quadruple, meaning that even the most powerful GPU available to us (A100, on Colab) cannot handle inputs as large as 6000 tokens. Given that T5 was originally trained on inputs of only 512 tokens, we did not anticipate effective summarization performance for our larger input sizes.
Applying the LORA technique to T5 could mitigate some memory issues by performing Single Value Decomposition on the training matrices, which simplifies the complexity of the matrices by retaining only their most critical components. However, this does not overcome the fundamental limitation of T5's original training on shorter sequences.
Reducing our input size to fit the model's capacity severely hampers its ability to summarize effectively. Since the average input size is approximately 7000 tokens, truncating to 1000 tokens means T5 misses

about 6000 tokens on average, which practically eliminates its capability to generate meaningful summaries.

BioBART

After setting the baseline, our approach centered on leveraging state-of-the-art natural language processing (NLP) techniques to address the challenges associated with manual review and summarization of medical literature. Recognizing the vast volume of data produced by medical studies and the need for efficient summarization methods, we opted to employ a deep learning-based approach, specifically utilizing the BioBART model—a variant of BART (Bidirectional and Auto-Regressive Transformers) pre-trained on biomedical text.

To fine-tune the BioBART model for medical review summarization, we followed a systematic methodology. Initially, we preprocessed the dataset, extracting abstracts from multiple medical documents and tokenizing them to prepare them for input into the model. Additionally, we generated corresponding summaries for these abstracts to serve as target labels during the training process. This preprocessing step ensured that the data was appropriately formatted and aligned with the model's input requirements. Subsequently, we utilized the Hugging Face library to facilitate model training and evaluation. We divided the preprocessed dataset into training and validation sets and configured the training pipeline using the Seq2SeqTrainer module. This module encapsulates the training process, incorporating essential components such as data collation, tokenizer configuration, and metric computation.

During model training, we fine-tuned the BioBART model on the training dataset, optimizing its parameters to minimize the loss function while maximizing summarization performance. To evaluate the model's effectiveness, we used the ROUGE metric.

LongformerEncoderDecoder (LED)

We considered MS2 referenced paper approach but ultimately chose the LongformerEncoderDecoder (LED) model, detailed in "Longformer: The Long-Document Transformer." The LED model architecture is built for processing long documents.

Unlike traditional transformers where the attention head size quadruples with input size, LED's attention head size increases linearly, allowing it to manage up to 16,000 tokens. The LED model, which comes in two variants—led-base-16384 and led-large-16384, initialized from bart-base and bart-large respectively—proved to be more suitable for our needs.

We fine-tuned the allenai/led-base-16386 model with 10 epochs with 8K token input. Additionally the allenai/led-large-16386 model with over 2 epochs with 8K token input.

Extractive-Abstractive Summarization with Base Model ("allenai/led-base-16384")

We applied TF-IDF (term frequency–inverse document frequency) for extractive summarization on individual abstracts, ranks sentences within each abstract by importance, and selects the top sentences to concatenate into a summary. These summaries had an average string size of 17,695 (ranging from 24 to 334,689) and an average token size of 2,951 (ranging from 4 to 54,810). These extractive summaries were then inputted into the LED model to create abstractive summaries. We experimented with two configurations of the LED model, using maximum input lengths of 3,000 and 6,000 tokens. The performance on the base model is not specified here.

BioClinic Tokenizer
Integrating "emilyalsentzer/Bio_ClinicalBERT" into our process resulted in a very high rouge-1 recall score (0.37), but also resulted in several inaccuracies and nonsensical terms. It did not provide minimum viable summaries. For brevity the actual sample is omitted.

Result and Discussion

The best two Summaries generated by our models are as follows.

(target) from our dataset: *"The use of glucomannan did not appear to significantly alter any other study endpoints.\nPediatric patients , patients receiving dietary modification , and patients with impaired glucose metabolism did not benefit from glucomannan to the same degree .\nGlucomannan appears to beneficially affect total cholesterol , LDL cholesterol , triglycerides , body weight , and FBG , but not HDL cholesterol or BP"*

- **Full Abstracts with 8K input tokens on large LED model ("allenai/led-large-16384")**
  For the larger version of the LED model F-measures for ROUGE-1 and 2 is 0.21973, 0.04915.
  Sample summary generated: *"The results of this meta- analysis suggest that glucomannan fibers may be effective in the treatment of hyperglycemia. However, there was no evidence for any effect on blood lipids. There was no significant difference between the two groups in the effects of the two types of dietary interventions"*
  Qualitative Review: This is an excellent review where enough details and technical terms are provided to generate confidence in the reader about the summary, without providing too many details that would convolute the message. It also includes positive results in the summary with the caveat of missing evidence.
- **Extractive-Abstractive Summarization with Large LED Model ("allenai/led-large-16384")**
  Rouge1 and Rouge2 scores with 6K Token input size were 0.229000, 0.046500 and with for 3K input it was 0.234000
  Sample summary generated for 3K: "*There was no evidence that glucomannan or any of the other interventions were more effective than placebo. There is limited evidence to support the use of fiber supplements in the treatment of overweight and obese adults."*
  Sample summary generated for 6K: "*There is a lack of evidence to support or refute the use of fiber supplements for the treatment of type 2 diabetes."*
  Qualitative Review: Both summaries state the correct conclusion, and for the 3K model, domain knowledge technical terms are included, however for the 6K model, the technical term is abstracted as fiber supplements; this negatively affects the Rouge scores.

Among all of the experiments we have done, we chose the large LED model, 'e_l_longformer_8k_plain1/checkpoint-2500/', as our top performer. Configured as the "allenai/led-large-16384" model, it handles full abstracts with an 8K token input size and has been fine-tuned over 2 epochs. This model achieved the highest ROUGE-2 score, along with a high ROUGE-1 score. ROUGE-2 specifically measures the overlap of bigrams (pairs of consecutive words) between the input data and the generated summary, indicating this model's superior ability to capture complex structures and relationships within the text compared to others.

The next most effective models were similarly large but fine-tuned with 6K token inputs and used extracted abstracts.

Hyperparameter search

We conducted a hyperparameter sweep to enhance the quality of outputs based on BERTScore. Below is a table summarizing the specific hyperparameters we adjusted, along with the rationale for each setting:

Hyperparameter	Values	Purpose
`num_beams`	range (2, 7)	Higher beam counts may yield more legible results
`length_penalty`	[1.0, 2.0, 3.0, 4.0]	Encourages the model to generate longer summaries.
`no_repeat_ngram_size`	range (2,5)	Settings of 3 or higher could improve ROUGE-2 scores.
`early_stopping`	[True, False]	Aims to speed up the generation process.

Table1: Hyperparameter Optimization for Enhanced Summarization Performance

We continue to refine the model's performance, we adjusted several hyperparameters and recorded the effects on ROUGE scores and summarization quality. The adjustments and their specific settings are detailed in the table below:

Hyperparameter	Values	Purpose
`num_beams`	range (5,7)	Narrowed range due to observed decrease in ROUGE scores with higher beams
`length_penalty`	[1.0, 1.25, 1.5]	Encourages the model to generate outputs in similar length to the target.
`no_repeat_ngram_size`	range (2,4)	Increased to above 2 to prevent poor quality results from repeated bi-grams
`early_stopping`	[False]	Chosen to forgo speed in favor of achieving the best results

Table 2: Hyperparameter Optimization for Enhanced Summarization Performance

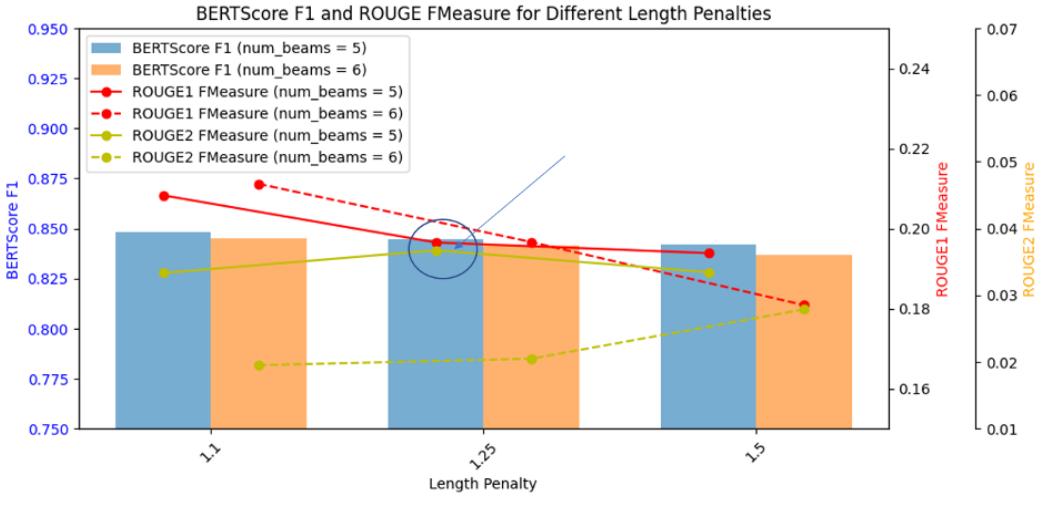

Figure 1: BERTScores and ROUGE Scores vs. Length_Penalty.

Figure 1 above shows that using 5 beams produces slightly better BERTscores compared to 6 beams. We also see that BERTScore F1 remains almost the same with a Length_Penalty of 1.25, but the

ROUGE2_FMeasure increases. Therefore, we choose this configuration for our final model.

`num_beams`	`length_penalty`	`no_repear_ngram_size`	`early_stopping`
5	1.25	2	False

Table 3: Final hyperparameters chosen for our model

Our BERTScore suggests that the summary is semantically like the abstracts and captures the essence well. However, the lower ROUGE scores for bigrams indicate that the exact wording and specific nuances may not align closely with the target.

Result	Rouge1	Rouge2	Fmeasure	BERTScore
MS2 Paper-LED Model	0.2689	0.0891	0.4500	N/A
Large LED Model	0.28118	0.057698	0.048079	[Precision: 0.8585, Recall: 0.8427, F1 Score: 0.8503]

Table 4: Final Results of our Tuned LED Model vs. the LED Model from the MS2 Paper

Conclusion: Further work and Challenges

We had challenges fine tuning the "allenai/led-large-16384" model, as it required a lot of resources. For future work, further fine tuning can be made.

The next best performing models are the same large models fine-tuned with 6K inputs but with extracted abstract inputs, further investigation into this approach can be made, as 6K inputs did not pose the fine-tuning challenges as with 8K. For future work a domain knowledge SME must be recruited to help us understand the biomedical terminology. Finally, another search for a tokenizer centered around biomedical data can be made.

A brand-new technique (March 2024) LoRA+ did not work as expected. It was expected to improve the performance but in most iterations it didn't help. In some models, it didn't work, and returned errors.

During the fine-tuning processes, we were restricted with the resources. We used commercially available virtual machines and managed ML platforms by cloud providers. The challenges of managing the costs restricted us for further development.

References:

DeYoung, Jay, et al. "MS2: Multi-Document Summarization of Medical Studies" arXiv e-prints, arXiv:2104.06486 (2021). https://arxiv.org/abs/2104.06486

Das, Mamata et al. "A Comparative Study on TF-IDF Feature Weighting Method and its Analysis using Unstructured Dataset." arXiv e-prints, arXiv:2308.04037 (2023). https://arxiv.org/abs/2308.04037

Yuan, Hongyi et al. "BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model ." arXiv e-prints, arXiv:2204.03905 (2022). https://arxiv.org/abs/2204.03905

Giorgi, John, et al. “Open Domain Multi-document Summarization: A Comprehensive Study of Model Brittleness under Retrieval” arXiv e-prints, arXiv:2212.10526 (2022) https://arxiv.org/abs/2212.10526

Lewis, Mik et al. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension” arXiv e-prints, arXiv:1910.13461 (2019) https://arxiv.org/abs/1910.13461

Hu, Edward J. et al. “LoRA: Low-Rank Adaptation of Large Language Models” arXiv e-prints, arXiv:2106.09685 (2021) https://arxiv.org/abs/2106.09685

Hayou, Soufiane et al. “LoRA+: Efficient Low Rank Adaptation of Large Models” arXiv e-prints, arXiv:2402.12354 (2024) https://arxiv.org/abs/2402.12354

Huggingface Model Links

https://huggingface.co/datasets/allenai/mslr2022
https://huggingface.co/docs/transformers/en/model_doc/t5#transformers.T5Model
https://huggingface.co/docs/transformers/en/model_doc/led
https://huggingface.co/allenai/led-base-16384
https://huggingface.co/allenai/led-large-16384
https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

Appendix:

Full Abstracts with 8000 input tokens on base LED model ("allenai/led-base-16384")
We fine-tuned the allenai/led-base-16386 model with 10 epochs with 8K token input. The result was of a ROUGUE-2 f-measure 0.0417.

(target) from our dataset: *"The use of glucomannan did not appear to significantly alter any other study endpoints.\nPediatric patients , patients receiving dietary modification , and patients with impaired glucose metabolism did not benefit from glucomannan to the same degree .\nGlucomannan appears to beneficially affect total cholesterol , LDL cholesterol , triglycerides , body weight , and FBG , but not HDL cholesterol or BP"*

Sample summary generated: *"There was no evidence of an effect of glucomannan on body weight, blood pressure, body mass index or body composition.\nCONCLUSIONS Glucoman may be an effective lipid-lowering intervention in children and adults, but there is insufficient evidence to recommend its use in adults"*

Qualitative Review: The summary is succinct and although it does not specify all the details it does provide an accurate overview summary. In addition, it includes domain knowledge technical terms. However, there are artifacts in the generated summary such as "n\CONCLUSIONS".

Extractive-Abstractive Summarization with base LED Model ("allenai/led-base-16384")

Sample summary generated for 3K: "*The results of this systematic review suggest that glucomannan is an effective treatment for hyperlipidemic type 2 diabetes."*

Sample summary generated for 6K: "*Conclusion: Glucomannan is an effective adjunct to the regular diet for the treatment of hyperlipidemic type 2 diabetes."*

Qualitative review: Both summaries state the wrong conclusion, since neither summary included the caveat that there is insufficient evidence to state that it's an effective treatment. In addition, the summaries do not include domain knowledge technical terms.

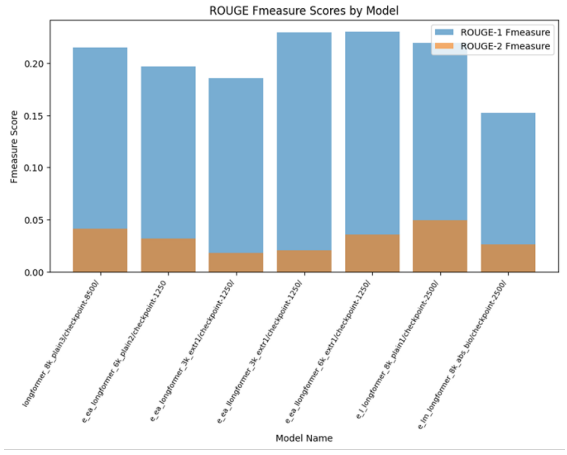

Figure 2: Rouge Fmeasure Scores by Model

Final Evaluation Results

Results			
Model	**ROUGE**	**Avg BLEURT Score**	**Qualitative Review**
Pegasus, pre-trained	'rouge1': 0.1739883726903796 'rouge2': 0.019690773116985447 'rougeL': 0.127040990614671 'rougeLsum': 0.1409497383867967	-0.690973102	N/A
T5, pre-trained	'rouge1': 0.17038833695850014 'rouge2': 0.0171524182758342 'rougeL': 0.10866151595889414 'rougeLsum': 0.1263354911004681	-0.992361688	N/A
T5, no config	'rouge1': 0.1399291833672796 'rouge2': 0.02022797674009731 'rougeL': 0.11192613991163428 'rougeLsum': 0.121634067979621	-1.063169876	N/A
T5, fine-tuned	'rouge1': 0.22742184304064095 'rouge2': 0.03907491763435844 'rougeL': 0.14340782037505348 'rougeLsum': 0.17390938234873882	-0.539650275	N/A
BioBART	'rouge1: 0.1546 'rouge2: 0.0288	N/A	N/A
BioBART_LORA	'rouge1: 0.161149 'rouge2: 0.028002	N/A	N/A
LED, base Full Abstracts 8k Input	'rouge2': 0.0411	N/A	Succinct, accurate overview with domain technical knowledge. Includes artifacts ('n\CONCLUSIONS')
LED, large Full Abstracts 8k Input	**'rouge1': 0.21973 'rouge2': 0.04915**	**N/A**	**Excellent, detailed review with technical terms. Includes positive results in summary with a caveat.**
LED, base Extractive-Abstractive 3k Input	'rouge1': 0.220300 'rouge2': 0.039200	N/A	Both summaries state wrong conclusion and do not include domain technical knowledge.
LED, base Extractive-Abstractive 6k Input	'rouge1': 0.244700 'rouge2': 0.045800	N/A	
LED, large Extractive-Abstractive 3k Input	'rouge1': 0.234000 'rouge2': 0.046500	N/A	Includes domain technical knowledge, however, technical terms are abstracted and negatively affect Rouge scores.
LED, large Extractive-Abstractive 6k Input	'rouge1': 0.229000 'rouge2': 0.046500	N/A	

Figure 3: Results