Transforming Healthcare with NLP: Automated Medical Report Summarization with Transformers

Subhrajit Dey[1], Dipanwita Rano[1], Akanksha Lohia[1], and Arindam Biswas[1]

[1]Department of Information Technology
Indian Institute of Engineering Science and Technology, Shibpur
India-711103
{2020itb099.subhrajit, 2020itb009.dipanwita,
2020itb073.akanksha}@students.iiests.ac.in, abiswas@it.iiests.ac.in

Abstract. Natural Language Processing (NLP) has witnessed significant advancements, particularly with Large Language Models (LLMs) like GPT series, revolutionizing text comprehension and generation. In healthcare, where managing vast amounts of medical data is crucial, LLMs offer promising solutions for summarizing complex medical reports efficiently. This paper explores the effectiveness of LLMs in automating medical report summarization, focusing on both extractive and abstractive approaches. Through fine-tuning models on the Cochrane dataset and leveraging state-of-the-art LLMs such as BART, Pegasus, and T5, we conducted a comparative analysis of summarization quality. Our methodology involved ensemble modeling and evaluating summaries using ROUGE metrics, augmented by TF-IDF and Paraphrase Mining scores for summary selection. Results indicate FLAN-T5 base consistently provided optimal summaries. However, resource constraints limited our study, highlighting the need for further exploration with larger models and datasets. Future research should address these limitations to advance medical report summarization, ultimately improving healthcare accessibility and outcomes.

Keywords: Transformers · Abstractive summarization · ROUGE score.

1 Introduction

In recent years, natural language processing (NLP) has seen incredible progress, especially with the introduction of Large Language Models (LLMs) like the GPT (Generative Pre-trained Transformer) series. These models have showcased extraordinary abilities in comprehending and producing text that is strikingly similar to human writing. Their prowess extends to various applications, but one of their most crucial uses is in text summarization. This task is all about condensing lengthy documents into shorter versions while ensuring that the most important details are preserved. LLMs tackle text summarization with finesse, employing sophisticated algorithms to sift through vast amounts of information and distill it down to its essence. Their ability to understand context and discern significance allows them to generate concise summaries that capture the essence of the original text. This has profound implications across numerous domains, from news articles and research papers to legal documents and medical reports.

In contemporary healthcare, the meticulous management of medical information stands as a paramount necessity. Physicians meticulously craft and navigate medical reports, comprehensive documents encapsulating a patient's health status. These reports encompass a spectrum ranging from the patient's ailment to treatment strategies and prognostic assessments. Their significance lies in furnishing doctors with insights necessary for informed decision-making regarding patient care. However, the copiousness of medical data poses a formidable challenge, wherein medical reports often manifest as protracted and intricate narratives akin to locating a needle in a haystack. This dilemma finds a potential solution in Natural Language Processing (NLP), a groundbreaking technology facilitating computers to comprehend human language.

Notably, Large Language Models (LLMs) represent a pinnacle in NLP, leveraging extensive text-based training to swiftly distill pivotal information from lengthy medical reports, akin to discerning salient points within an expansive narrative. This transformative capability markedly enhances efficiency, equipping physicians with rapid access to crucial insights without navigating through voluminous text. Hence, this study delves into the paradigm shift catalyzed by LLMs in medical documentation, particularly focusing on the automation of report summarization. Two primary methodologies, namely extractive and abstractive summarization, are explored, with the latter being the subject of investigation using the Cochrane HuggingFace dataset. Fine-tuning of several models, including the BART-Base model, Pegasus, and Google's T5-small model, was conducted on the Cochrane dataset, with emphasis on optimizing hyperparameters to enhance performance. The inherently intricate nature of the data necessitated meticulous preprocessing to facilitate summary generation. Comparative analysis, utilizing datasets provided by AllenAI, involved evaluating the efficacy of these models through the ROUGE metric, thereby elucidating the accuracy of generated summaries.

The primary contribution of this paper lies in the introduction of an ensemble learning approach featuring four distinct large language model architectures. These models collectively generate condensed summaries, which undergo a comparative evaluation using TF-IDF and paraphrase mining techniques. The most informative summary is subsequently selected based on this comparative analysis, thus enhancing the quality and informativeness of the generated summaries.

2 Background Study

We explored the various Text Summarization Techniques and classified text summarization into two broad categories namely, Unsupervised Summarization and Supervised Summarization.

Unsupervised summarization approaches include auto-encoders to mirror the information compression inherent in summarization [2][8][4] as well as large-scale pre-training for domain specific adaptation [21]. However, little work has focused on domain adaptation in summarization. Wang *et al* [20] examines domain adaptation for extractive summarization. Hua *et al.* [12] showed that summarization models have difficulty generating text in the style of the target domain, while more recently, Zhang *et al.*[22] report strong performance of pre-trained models when trained in few-shot settings and Bražinskas *et al.* [5] fine-tune dataset specific components of a model for few-shot learning. Fabbri *et al.* [9] uses pre-trained models and improves zero-shot and few-shot summarization by encoding characteristics of the target summarization dataset in unsupervised, intermediate fine-tuning data.

Previous studies [10],[13] focused on how to better utilize and digest information from EHRs to enhance the efficiency of healthcare services. Summarization is one of those techniques that can be applied, which generally has two approaches: abstractive and extractive summarization. As abstractive summarization sometimes fails to capture factual details accurately as needed in medical settings, extractive summarization was considered to be more suitable. This method directly extracts a subset of data written by the medical experts as the summary. Unsupervised extractive summarization was first explored [3],[11],[14]. Due to the recent success of neural networks, supervised approaches become more popular for extractive summarization [1],[6],[7]. One obstacle for training a supervised model for extractive summarization in the medical domain, however, is the lack of labeled data, since annotations for EHRs require disease-specific medical background and can be very expensive. Liu *et al.* [17] trains a supervised model, and studies how to utilize the intrinsic correlation between multiple notes for a single patient to generate pseudo-labels and guide summarization. Liang *et al.* [15] is targeted toward a single clinical encounter represented by a note and specific to the management of a given disease: hypertension and diabetes mellitus. It focuses entirely on extracting informative sentences rather than cohesive sentences from a single clinical note. Vinod *et al.* [19] works on the idea to expand BERTSUMEXT's knowledge to give it a 'medical edge' that it lacks by further training the BERTSUMEXT model using different training strategies on a clinical report summarization dataset. Evaluating the quality and effectiveness of medical report summarization systems requires appropriate metrics and benchmark datasets. Researchers have proposed various evaluation metrics, including ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [16], BLEU (Bilingual Evaluation Understudy) [18], and domain-specific metrics tailored to medical summarization tasks.

3 Proposed Methodology

An exhaustive study of various text summarization methodologies and the need for clinical report summarization led to the development of an architecture that meets the requirement of the system. Our proposed architecture is composed of an ensemble of various fine-tuned models for generation of medically informed condensed summaries. The architecture can be divided into 3 broad segments, namely, Preprocessing phase, Training phase, and Evaluation phase.

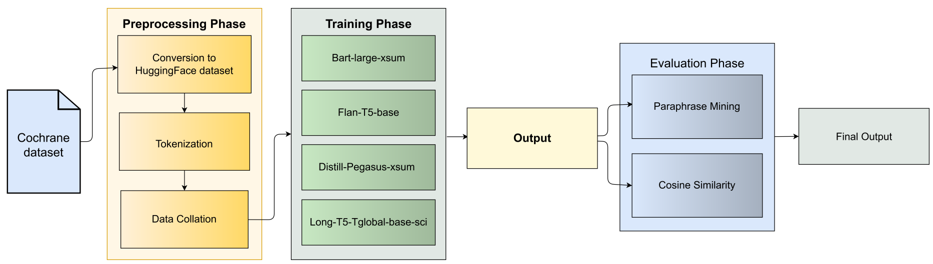

Fig. 1. Workflow of the automatic medical text summarizer and the evaluation metrics

3.1 Preprocessing phase

The preprocessing phase of our architecture begins with fetching abstracts from the Cochrane dataset, a collection of high-quality systematic reviews and meta-analyses in healthcare, which is followed by converting it in a form that can be used for fine-tuning our ensemble architecture. The preprocessing phase can be divided into 3 phases.

In the initial phase of preprocessing, our approach involves retrieving abstracts from the Cochrane dataset. Known for its meticulous systematic reviews, the Cochrane dataset endeavors to meticulously identify, evaluate, and amalgamate all empirical evidence meeting predefined eligibility criteria, all in pursuit of answering targeted research inquiries. Serving as a cornerstone of reliable information, it stands as an indispensable resource for care providers, recipients, decision-makers, and the academic community alike. Thus, it serves as a foundational element within our architectural framework, elevating the standards of our approach through its unparalleled depth and rigor.

The next phase of preprocessing involves tokenization. In order to achieve consistency, the length of abstracts have been standardized to 1024 tokens, and summaries to 400 tokens. Subsequently, specialized tokenizers have been tailored to our ensemble architecture to generate tokens of required lengths. The final phase of preprocessing involves Data Collation whereby the data is transformed to a form that can be used as inputs to the ensemble architecture. The summaries associated with multiple PMIDs were considered and merged to form a single summary, thus preparing our dataset for use in subsequent steps.

3.2 Training Phase

The pivotal stage of our architecture resides within the training phase, which comprises an ensemble model consisting of four large language models (LLMs) fine-tuned on the input dataset. Each LLM within this ensemble is tasked with generating a summary, which is subsequently forwarded to the subsequent layer for the generation of the final outcome. Specifically, the LLM models employed encompass Flan-T5-base, distill-pegasus-xsum, bart-large-xsum, and long-T5-global-base-sci, each contributing to the comprehensive synthesis of information.

Flan T5 base FLAN (Fused Local and Non-local features) is an innovation from Microsoft Research that boosts the power of T5 (Text-To-Text Transfer Transformer). FLAN-T5 skillfully blends local and non-local self-attention mechanisms. This combination allows the model to see close-knit connections within the text along with broader connections across the entire text. By tapping into non-local features, FLAN-T5 deepens the model's grasp of the context, resulting in summaries of exceptional quality. This approach is a breakthrough in transformer-based models, delivering greatly improved summarization capabilities by seamlessly combining local and global features.

Distill PEGASUS xsum This model is an advancement in text summarization, leverages transformer-based models. It stands out from traditional models by employing self-attention and a unique pre-training process called Gap Sentence Generation (GSG). GSG involves training PEGASUS on a massive text dataset where sentences are masked, and the model predicts the missing sentence. This pre-training allows PEGASUS to comprehend the text's context and generate accurate summaries. PEGASUS-xsum is a specialized version of the PEGASUS language model, designed to excel on the XSum dataset. XSum features brief news articles, making it a suitable testing ground for summarization models. By fine-tuning PEGASUS on XSum, its capability to create concise and meaningful summaries is improved, solidifying its position as a top abstractive summarization model.

BART Large xsum Bidirectional and Auto-Regressive Transformers developed by Facebook AI, is a transformer-based language model for sequence-to-sequence tasks. A specialized version, BART-large-xsum, has been trained on the XSum dataset, a collection of news articles and single-sentence summaries. This training enables BART-large-xsum to effectively summarize news articles, generating brief and informative summaries that match human-written examples. Leveraging its bidirectional and auto-regressive capabilities, BART-large-xsum adeptly captures pertinent information from input documents, enabling the generation of coherent and fluent summaries. Thus, within the realm of abstractive summarization tasks, particularly in the news domain, BART-large-xsum stands as a formidable tool.

Long T5 Tglobal base sci This model is a modified T5 (Text-To-Text Transfer Transformer) architecture. It's made for handling scientific text tasks. Thee model has a longer architecture designed for longer input sequences. This helps it understand complex scientific ideas and context better. With the "TGlobal" part, it focuses on understanding the overall context. This is important for summarizing challenging scientific literature well. Built on T5's flexibility, Long-T5-TGlobal-Base Sci is good for various scientific text applications. These include summarization and information extraction. Its base architecture balances efficient computing with high performance. This makes it useful for researchers and professionals in the scientific field.

We have fine-tuned the four above-mentioned models on the Cochrane dataset. With fine-tuning the models, we ensure that the models not only perform summarization, but is made domain-aware, and is capable of generating medical summarization. The models underwent a training regimen of 10 epochs, with a process involving saving and reloading the model after each epoch continuously. The training for the models was carried out using the HuggingFace Trainer Class.The training process employed the Seq2SeqTrainer and Seq2SeqTrainingArguments, both of which inherit from the Trainer and TrainingArgument classes, respectively. The Trainer comprises of comprehensive training and eval loop for PyTorch, which is optimised for Huggingface Transformers.

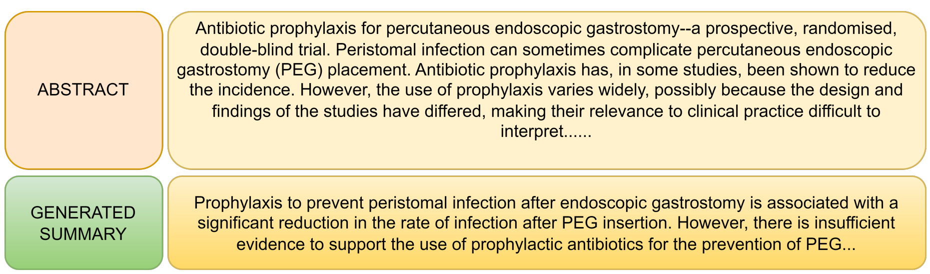

Fig. 2. Sample summary generated by the model

3.3 Evaluation Phase

The Evaluation phase, the concluding segment of our architecture, plays a crucial role in selecting the most informative summary produced by the ensemble architecture. This phase involves passing the summaries generated by all four models through Paraphrase Mining and TF-IDF to compute text overlap, with the resulting average serving as the score. Subsequently, the summary with the highest score is designated as the final output from the model. Further elaboration on the evaluation models is provided below.

Paraphrase Mining Paraphrase mining, crucial in numerous Natural Language Processing tasks such as Language Translation and Text Summarization, involves identifying sentences within a corpus that convey the same meaning but in different expressions. This paper makes use of paraphrase mining, capable of scoring and ranking all pairs within large corpora. However, it suffers from a significant limitation due to its quadratic runtime, rendering it inefficient for extensive datasets. To address this, we utilized the SentenceTransformer library with the "all-MiniLM-L6-v2" model to conduct paraphrase mining. By inputting summaries generated from the ensemble architecture of four transformer models, we established similarity relationships between each summary, ultimately selecting the one with the highest overall relationship as the final output.

TF-IDF Vectorization and Cosine Similarity To assess the relationship among summaries generated by the transformer models, we employed TF-IDF Vectorization followed by Cosine Similarity computation. TF-IDF, a fundamental NLP technique, evaluates word importance by comparing its frequency within a document to its occurrence across the corpus. This vectorization method transforms textual data into numerical vectors, aiding in quantifying textual relationships. Cosine Similarity, utilized to measure likeness between vector pairs in high-dimensional space, was applied to assess the similarity between summaries. The resultant relationship values facilitated the selection of the summary with the highest overall relationship as the final output, enhancing the accuracy and informativeness of the generated summaries.

4 Results

4.1 ROUGE Scores

We list down the performance metrics of the fine-tuned models, including ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum scores. Table 1 summarizes the ROUGE scores for each of the four finetuned models evaluated in our study. The ROUGE-1 scores indicate the overlap of unigrams between the generated summaries and the reference summaries, while ROUGE-2 measures the overlap of bigrams. ROUGE-L considers the longest common subsequence between the summaries and the references, whereas ROUGE-Lsum computes the F1 score based on the longest common subsequence divided by the total number of words in the reference summary.

4.2 Summary Comparison Results

The average score from "Paraphrase Mining" and "TF-IDF Vectorization and Cosine Similarity" were taken for each summary and the final summary was selected for each abstract. 250 abstracts from the test dataset were selected and the summaries generated by the ensemble architecture was compared among themselves. The summary with the highest similarity with the summaries generated

Table 1. ROUGE score comparisons of the finetuned models

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
Flan T5 base	0.2640	0.0367	0.1687	0.1699
Distll PEGASUS xsum	0.2212	0.0341	0.1600	0.1605
BART Large xsum	0.2159	0.0274	0.1418	0.1431
Long T5 Tglobal base sci	0.1958	0.0255	0.1359	0.1368

by the other transformer models was selected. The results have been provided in the following table.

Table 2. Number of summaries selected as the final output for each model on the basis of Paraphrase Mining scores and Cosine Similarity scores individually and finally with the average of the two scores combined together

Model	Paraphrase Mining	TF-IDF and Cosine Similarity	Average Score
Flan T5 base	126	183	157
Distll PEGASUS xsum	68	52	69
BART Large xsum	7	1	1
Long T5 Tglobal base sci	49	14	23

This is clear from the results given above, that the most ROUGE score was assigned to fine-tuned FLAN-T5 Base model among all the models tuned in this paper. Also, this model has contributed much more than others to the final output with 157 out of 250 summaries in our test dataset were selected from its output. From the above findings, considering other evaluated models, it can be concluded that the fine-tuned FLAN-T5 Base model has performed excellently in medical report summarization.

5 Conclusion

In our investigation, we explored the application of NLP techniques to streamline medical reports, leveraging four refined models: BART-large-xsum, FLAN-T5, PEGASUS-xsum, and Long-T5-TGlobal-Base Sci. Through extensive experimentation, we devised succinct summaries for medical reports using our proposed methodology, ultimately selecting the most effective summary based on the highest evaluation score. Our findings consistently highlighted FLAN-T5 base as the top performer, as evidenced by Table 2, which illustrates its superiority in generating optimal summaries. However, our study encountered limitations, including restricted GPU support and storage space constraints, leading to reduced dataset usage and truncated model training epochs.

Looking ahead, future investigations in this domain could benefit from employing more expansive and sophisticated models such as BART-large or PEGASUS-large, potentially enhancing the quality of summarization outcomes. By addressing the challenges encountered in this study and delving deeper into the capabilities of advanced NLP models, opportunities emerge for the development of more precise and comprehensive medical report summarization systems. This advancement aims to improve access to critical medical information and elevate healthcare outcomes.

References

1. Arumae, K., Liu, F.: Reinforced extractive summarization with question-focused rewards. arXiv preprint arXiv:1805.10392 (2018)
2. Baziotis, C., Androutsopoulos, I., Konstas, I., Potamianos, A.: Seqˆ 3: differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. arXiv preprint arXiv:1904.03651 (2019)
3. Brandow, R., Mitze, K., Rau, L.F.: Automatic condensation of electronic publications by sentence selection. Information Processing & Management **31**(5), 675–685 (1995)
4. Bražinskas, A., Lapata, M., Titov, I.: Unsupervised opinion summarization as copycat-review generation. arXiv preprint arXiv:1911.02247 (2019)
5. Bražinskas, A., Lapata, M., Titov, I.: Few-shot learning for opinion summarization. arXiv preprint arXiv:2004.14884 (2020)
6. Cao, Z., Li, W., Li, S., Wei, F., Li, Y.: Attsum: Joint learning of focusing and summarization with neural attention. arXiv preprint arXiv:1604.00125 (2016)
7. Cheng, J., Lapata, M.: Neural summarization by extracting sentences and words. arXiv preprint arXiv:1603.07252 (2016)
8. Chu, E., Liu, P.: Meansum: a neural model for unsupervised multi-document abstractive summarization. In: International conference on machine learning. pp. 1223–1232. PMLR (2019)
9. Fabbri, A.R., Han, S., Li, H., Li, H., Ghazvininejad, M., Joty, S., Radev, D., Mehdad, Y.: Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. arXiv preprint arXiv:2010.12836 (2020)
10. Ford, E., Carroll, J.A., Smith, H.E., Scott, D., Cassell, J.A.: Extracting information from the text of electronic medical records to improve case detection: a systematic review. Journal of the American Medical Informatics Association **23**(5), 1007–1015 (2016)
11. He, Z., Chen, C., Bu, J., Wang, C., Zhang, L., Cai, D., He, X.: Document summarization based on data reconstruction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 26, pp. 620–626 (2012)
12. Hua, X., Wang, L.: A pilot study of domain adaptation effect for neural abstractive summarization. arXiv preprint arXiv:1707.07062 (2017)
13. Lee, J.Y., Park, H.A., Chung, E.: Use of electronic critical care flow sheet data to predict unplanned extubation in icus. International Journal of Medical Informatics **117**, 6–12 (2018)
14. Li, C., Qian, X., Liu, Y.: Using supervised bigram-based ilp for extractive summarization. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1004–1013 (2013)
15. Liang, J., Tsou, C.H., Poddar, A.: A novel system for extractive clinical note summarization using ehr data. In: Proceedings of the 2nd clinical natural language processing workshop. pp. 46–54 (2019)

16. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013
17. Liu, X., Xu, K., Xie, P., Xing, E.: Unsupervised pseudo-labeling for extractive summarization on electronic health records. arXiv preprint arXiv:1811.08040 (2018)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). https://doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040
19. Vinod, P., Safar, S., Mathew, D., Venugopal, P., Joly, L.M., George, J.: Fine-tuning the bertsumext model for clinical report summarization. In: 2020 International Conference for Emerging Technology (INCET). pp. 1–7. IEEE (2020)
20. Wang, D., Liu, P., Zhong, M., Fu, J., Qiu, X., Huang, X.: Exploring domain shift in extractive text summarization. arXiv preprint arXiv:1908.11664 (2019)
21. Yang, Z., Zhu, C., Gmyr, R., Zeng, M., Huang, X., Darve, E.: Ted: A pretrained unsupervised summarization model with theme modeling and denoising. arXiv preprint arXiv:2001.00725 (2020)
22. Zhang, J., Zhao, Y., Saleh, M., Liu, P.: Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In: International conference on machine learning. pp. 11328–11339. PMLR (2020)

6 Appendix

Abstract	Antibiotic prophylaxis for percutaneous endoscopic gastrostomy--a prospective, randomized, double-blind trial. Peristomal infection can sometimes complicate percutaneous endoscopic gastrostomy (PEG) placement. Antibiotic prophylaxis has, in some studies, been shown to reduce the incidence. However, the use of prophylaxis varies widely, possibly because the design and findings of the studies have differed, making their relevance to clinical practice difficult to interpret. To determine the efficacy of antibiotics, either prophylaxis or concurrent antibiotics at the time of the procedure, in reducing peristomal infection after PEG insertion in the context of a study designed to reflect current practice. One hundred and forty-one patients undergoing PEG placement were randomized to group one to receive either a single dose of 750 mg of intravenous cefuroxime (n=50) or placebo (n=51) 30 minutes before PEG insertion. Forty patients who, for various reasons, were already receiving antibiotics were allocated to group two. The peristomal site was evaluated on days 3, 5, and 7 following insertion. Erythema and exudate were scored on a scale from 0 to 4; induration was scored on a scale from 0 to 3. A maximum combined score of 8 or higher or the presence of pus was criteria for infection. The primary outcome measure was the occurrence of a peristomal wound infection at any time within one week of PEG insertion. Peristomal wound infection was significantly reduced in patients who received antibiotics either as a single dose of cefuroxime [one of 33 (3percent)], or in those on antibiotics for prior indications [one of 36 (3percent)], compared with placebo [six of 33 (18percent)], P=0.04 and 0.03, respectively. Antibiotics, either prophylaxis or concurrent, reduce the incidence of peristomal wound infection after PEG placement.
Reference Summary	Administration of systemic prophylactic antibiotics for PEG tube placement reduces peristomal infection.

Summary Generated by BART-Large-xsum	Based on the results of this systematic review the administration of antibiotic prophylaxis for elective endoscopic gastrostomy cannot be universally recommended. Neither can the administration be recommended against when high rates of wound infection are observed.
Summary Generated by Flan-T5-base	Prophylaxis to prevent peristomal infection after endoscopic gastrostomy is associated with a significant reduction in the rate of infection after PEG insertion. However, there is insufficient evidence to support the use of prophylactic antibiotics for the prevention of PEG.
Summary Generated by Distill-Pegasus-xsum	Antibiotic prophylaxis reduces the incidence of peristomal infection after percutaneous endoscopic gastrostomy.
Summary Generated by Long-T5-Tglobal-base-sci	There is insufficient evidence to draw firm conclusions regarding the effectiveness of antibiotic prophylactic treatment for PEG.
Selected Final Summary	Prophylaxis to prevent peristomal infection after endoscopic gastrostomy is associated with a significant reduction in the rate of infection after PEG insertion. However, there is insufficient evidence to support the use of prophylactic antibiotics for the prevention of PEG.(**MODEL - Flan-T5-base)**

Fig. 3. Example Abstract and the Respective Summaries generated by the Transformer models