Biomedical Literature Review Summaries Using Multi-Document Summarization

Farid Gholitabar and **Mike Khor** and **Amir Moazami**
farid.gholitabar@berkeley.edu mike.khor@berkeley.edu amirmoazami@berkeley.edu

University of California Berkeley

Abstract

This research leverages multi-document summarization (MDS) to enhance literature review methods in the biomedical domain, aiming to generate concise, informative abstracts from diverse biomedical literature. We address challenges such as reconciling conflicting findings, managing complex terminology, and handling large datasets. Our approach integrates new and advanced NLP analysis techniques for deeper understanding than traditional methods. A hybrid extractive-abstractive model, incorporating innovative techniques tailored to biomedical literature, effectively condenses information into comprehensive summaries. Our findings highlight the high performance of the PEFT BioBERT + K-means + LongT5 model for short target sentences, suggesting promising avenues for future research. This work represents a significant step towards more efficient literature review methodologies and showcases the transformative potential of MDS in biomedical research.

1 Introduction

The exponential growth of biomedical research literature presents a formidable challenge for researchers to stay abreast of the latest developments and findings. Conducting literature reviews is a fundamental component in the biomedical domain and essential for assessing the efficacy and safety of medical interventions. However, this process is highly time-consuming, often taking approximately one to two years to review and synthesize information from individual medical studies thoroughly. The necessity for efficient and accurate methods to expedite this process is more pressing than ever.

Our project aims to address this challenge by applying multi-document summarization (MDS) techniques. Specifically, we propose to use MDS to create concise, coherent abstracts of literature reviews for biomedical journal topics. We aim to streamline the summarization of multiple related medical studies and their subsequent updates by employing existing algorithms and datasets.

Key challenges include: **i) conflicting findings**: varying and sometimes conflicting results in biomedical studies; **ii) complex vocabulary**: specialized lexicon of the biomedical field necessitating a deep understanding; **iii) long sequences**: handling a variable number of input studies, sometimes exceeding 20, requires robust data engineering and processing strategies.

Our research aims to understand and tackle these challenges and provide an MDS solution that enhances literature review processes. We hope to aid medical professionals and researchers in making informed decisions based on a short yet comprehensive summary of existing medical literature.

2 Background

2.1 Related Work and Previous Research

Popular summarization datasets (i.e., Multi-XScience [1]) are not specific to biomedical texts, whereas the "Cochrane" biomedical summarization dataset [2] lacks in size. Recently, the "Multi-Document Summarization of Medical Studies," MS^2 dataset [3] emerged, containing over 470,000 documents and 20,000 summaries. This large-scale public dataset facilitates automated biomedical MDS development and has sparked research efforts ([4, 5, 6]).

MDS techniques are diverse, encompassing a range of methods from graph-based approaches that leverage textual connections to context-based strategies that integrate broader meanings. For an insightful overview of these techniques, Ma et al. ([7]) provide comprehensive coverage. Foundational models like Bi-directional and Auto-Regressive Transformers (BART) have been instrumental in advancing abstractive summarization methods. However, the exploration of hybrid multistage extractive-abstractive approaches,

which combine the strengths of both extractive and abstractive methods, has been relatively limited. Such hybrid methods typically utilize an extractive step to condense input lengths, followed by an abstractive model to generate the final summary output. A notable example of this technique is WikiSum ([8]), which employs an initial extractive process to identify key information and then generates a summary based on these highlights. To date, Shinde et al. ([6]) are among the few who have investigated the application of multistage summarization in biomedical literature, specifically using the Cochrane dataset. Their approach combines a BERT-based extractive method with PEGASUS for abstractive summarization.

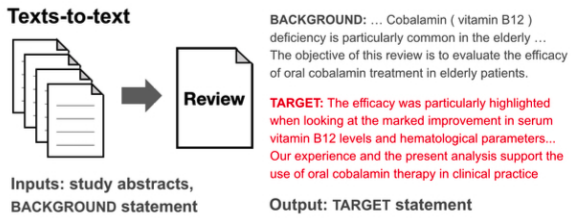

Figure 1: Given study abstracts and a BACKGROUND statement, generate the TARGET summary. Diagram from [3].

2.2 How is Our Work Different?

In our research, we introduce extractive-abstractive summarization tailored for the MS^2 dataset and offer a detailed comparative analysis of the choices of the extractive and abstractive components. By selectively parsing critical data from diverse abstracts using extractive summarization, our method overcomes the input length limitations of pure abstractive models. This refined approach enhances the quality of abstractive model output summaries.

In addition, we present several finetuning efforts on the chosen abstractive model. We highlight some important considerations for finetuning and detail the limitations of finetuned summaries.

3 Methods

3.1 Data Acquisition and Preprocessing

MS^2 contains 20K biomedical literature reviews (examples), and each review contains a background on the topic of interest, a list of studies referenced in the review, the studies' abstracts, and a target summary (text extracted from the "abstract" portion of the review) (Figure 1). We used the dataset's pre-split of training and validation sets but excluded the test pre-split because it lacks targets (used for leaderboard benchmarking). Prior to extractive summarization, MS^2 studies' abstracts are separated into sentences using the Scispacy module, a model trained explicitly on scientific literature.

3.2 Baseline Models

We employed extractive and abstractive models separately before progressing to a hybrid model to assess the individual strengths of each model type in isolation.

We assessed BERT[9], BioBERT[10], and SciBERT[11] models for extractive summarization, which provide a semantic understanding of individual sentences. We used K-means to cluster sentence-level BERT embeddings per study to pinpoint the most pertinent sentences and phrases. Then, we picked one representative sentence from each cluster, as inspired by [12]. This extractive technique retrieves sentences of diverse semantic meanings while leaving out redundant sentences. The choice of BERT encoders depends on their ability to understand complex biomedical topics and catch nuanced details.

We selected Pegasus[13], BioBART[14], and LongT5[15] models for abstractive summarization, which are proficient in generating coherent abstractive summaries. These models have constraints, such as Pegasus and BioBART's restricted input token length, which can hinder their ability to process longer documents. BioBART, while tailored for biomedical texts, could struggle with highly specialized or rare medical terminology not covered in its training dataset.

Please see Appendix A for a more detailed description of the abovementioned models.

3.3 Evaluation Methods

Our study utilized a specific segment of our dataset to evaluate model effectiveness in summarization tasks. The metrics employed were **ROUGE**, **BERTScore**, and **ΔEI** (Evidence Inference Delta). Specifically, ΔEI, as delineated in MS^2 [3], utilizes a Biomed-RoBERTa-based model to calculate probabilities across three evidence directions: "significantly decreases," "no significant difference," or "significantly increases." This metric evaluates the divergence in these probabilities between the target text and our generated summaries. A lower ΔEI score signifies greater agreement in the direction of evidence, reflecting the accuracy of our summarization models in capturing evidence directions.

3.4 Experimental Methods

Our research explored two distinct experimental methods, each employing a hybrid of extractive and abstractive techniques but with different approaches to either component. Each outlined experiment's architecture overview can be found in Appendix B.

3.4.1 Experiment 1: Hybrid Extractive-Abstractive Technique Using BioBERT and K-means Clustering with LongT5

Leveraging the synergy between BioBERT's advanced capabilities and K-means clustering for processing biomedical literature, our research employed a hybrid extractive-abstractive technique. This extracted text segment was then subjected to the Long T5 model.

We utilized K-means clustering to identify and select the most informative sentences in each abstract. We could isolate key sentences in each abstract by fitting the sentence embeddings into a K-means model configured with a specific number of clusters to match our desired sentence count. A unique aspect of our approach was ensuring that each chosen sentence came from a different cluster, thereby enhancing the diversity and relevance of our extracted content.

In the next phase, we concatenated the selected sentences to form extractive summaries and then fed them into LongT5, resulting in succinct abstractive summaries.

3.4.2 Experiment 2: Text Clustering and Summarization using BioBERT, UMAP, HDBSCAN, and LongT5

First, we used BioBERT to generate sentence embeddings, consistent with our previous approach. Next, we explore a different clustering technique for extracting key sentences, using Hierarchical Density-based Spatial Clustering of Applications with Noise (HDBSCAN)[16], which handles noise well and efficiently manages different data densities.

However, HDBSCAN struggles with high-dimensional data. To overcome this, we incorporated Uniform Manifold Approximation and Projection (UMAP) for dimension reduction, which retains density information, a critical aspect often lost in other dimension reduction methods.

Like K-Means extraction, one key sentence is extracted from each cluster of sentences and concatenated before being fed into LongT5.

3.4.3 Experiment 3: Rule-based Extractive Summarization Using Targeted Keyword Search

As a more straightforward extractive approach, our "extractive model" in this experiment was designed to search specifically for the word "conclusion" within the texts for each abstract. Upon locating this term, the model extracted the sentence containing it and any subsequent sentences in the abstract. This approach was based on the hypothesis that sentences following the term "conclusion" in scientific literature are likely to contain key summary points of the research.

After extracting these segments, we feed them into LongT5 to synthesize a cohesive summary. This method leverages the inherent structure of scientific abstracts and mimics how a human reviewer would approach this summarization problem.

3.4.4 Experiment 4: Extractive-Abstractive Summarization with BioBERT, K-means Clustering, and Fine-Tuned LongT5

In this experiment, we employed a two-phase training process for the LongT5 model, with each phase addressing different data representations.

During the initial phase, we trained the LongT5 model on the raw training data from the MS^2 dataset. This training, executed over 8 epochs, was strategically focused on minimizing cross-entropy loss. A notable constraint encountered was the model's limited input capacity, compounded by our computational resource limitations. Consequently, we were restricted to input data comprising 2,048 tokens for each training instance. However, this experiment led us to explore an alternative training approach in the second phase.

The second phase involved training the LongT5 model on a transformed version of the training data. We first processed the data through BioBERT for embedding, followed by the K-means, which provided a compressed and more focused representation of the input data. (how it was explained earlier in other experiments) This approach aimed to encapsulate the essential information effectively within the model's input capacity limits. We hypothesized that this representation would enable more comprehensive learning from the dataset. This phase of training also spanned 7 epochs.

To rigorously evaluate the model's performance

in each phase, we utilized two distinct sets of validation data. The first set entailed assessing the summarization capabilities of each model iteration using the unprocessed validation dataset. In contrast, the second set involved validation data that had undergone preprocessing through BioBERT embedding, followed by K-means clustering. This approach, aimed at creating a condensed and refined version of the input data, is indicative of a potential operational data pipeline for the model's inferences. Our dual-method testing approach was instrumental in thoroughly evaluating the model's proficiency in processing and summarizing both raw and pre-processed biomedical texts.

3.4.5 Experiment 5: Extractive-Abstractive Summarization with BioBERT, K-means Clustering, and PEFT LongT5

Lastly, we used a Parameter Efficient Fine-Tuning (PEFT) method, Low-Rank Adaptation (LoRA) to finetune *BioBERT-K-means-LongT5*. LoRA significantly reduces the number of parameters needed to fine-tune a pretrained model, drastically reducing memory usage during training [17]. The motivation for using LoRA stems from memory constraints encountered in full fine-tuning (the last experiment), in which the input tokens had to be truncated down from 16,384 to 2,048. With LoRA, we were able to fine-tune LongT5 with a longer input length of 8,192 using the same hardware. This longer input length means that only 5.3% of our BioBERT-K-means extractive summaries were truncated prior to LongT5 abstractive summarization.

PEFT LoRA was applied to all query and key weights for all 11 transformer layers (both encoder and decoder). Keeping our LongT5's original 247 million parameters frozen, we added an additional 1.7 million LoRA parameters, which means only 0.7% of parameters are trainable. The PEFT model is then trained on the training dataset for 4 epochs.

4 Results and Discussion

4.1 Choosing Appropriate Abstractive and Extractive Models

For abstractive summarization, this research evaluated Pegasus, BioBART, and LongT5 (Table 2). LongT5 is chosen due to its higher BERTscore average F1 and ROUGE-1 F1 scores, which indicate better performance in generating coherent, context-rich summaries. A significant advantage of LongT5 is its ability to process longer texts, which is crucial for handling extensive biomedical literature without needing data engineering modifications like restricting attention mechanisms. Its mixed attention mechanism, blending local and sparse global attention, is noted for efficiently handling large datasets, making it suitable for extensive text summarization tasks.

For extractive summarization, we selected BioBERT combined with K-means clustering for its low ΔEI average score of .503. This score reflects BioBERT's superior factual accuracy and alignment with target summaries. BioBERT's training in biomedical literature is a key factor in its performance, as it enables a nuanced understanding of biomedical contexts. This aspect is particularly valuable for information extraction from complex biomedical texts.

4.2 Evaluating the Combination of Extractive and Abstractive Models

We assessed three unique combinations of integrating extractive and abstractive methods: 1) *Rule-based + LongT5*, 2) *BioBERT + K-means + LongT5*, and 3) *BioBERT + UMAP + HDBSCAN + LongT5*. Notably, these combined models outperformed standalone BioBERT-K-means or LongT5, highlighting the complementary strengths of both approaches. The K-means algorithm effectively reduced input lengths for LongT5, improving attention management. LongT5 then condenses content into more precise outputs, thereby improving ROUGE precision.

ROUGE score analysis highlighted that the *BioBERT + UMAP + HDBSCAN + LongT5* model underperformed, often paraphrasing content and omitting key phrases or terms. In contrast, BERTScore evaluation, which focuses on semantic similarity, identified *Rule-based + LongT5* as more effective at capturing the essential semantic content despite lower ROUGE scores.

The ΔEI scores pointed to moderate inconsistencies in factual accuracy across all models compared to the pure extractive method. Of the three combinations, *BioBERT + K-means + LongT5* scored the lowest (best) ΔEI (.519) and preserved factual details better than the pure abstractive method (ΔEI = .543), as a contribution of BioBERT-K-Means. The model's effectiveness likely stems from the synergistic combination of BioBERT's understanding of biomedical contexts, K-means clustering efficiency, and LongT5 summarization. This blend facilitates

Experiment	Model	R-1	R-2	R-L	BERT score	ΔEI average	ΔEI macro F1
Baseline	LongT5	.185	.022	.106	.543	.543	.364
Baseline	BioBERT + K means	.046	.012	.029	.502	.503	**.389**
1	BioBERT + K means - LongT5	.194	.024	.109	.550	.519	.382
2	BioBERT + UMAP + HDBSCAN + LongT5	.187	.023	.105	.545	.559	.363
3	Rule-based (conclusion-focused) + LongT5	**.210**	**.028**	**.125**	.567	.526	.377
4	Full Fine-Tuned LongT5	.154	.022	.115	.561	.576	.332
4	BioBERT + K means - Full Fine-Tuned LongT5	.148	.018	.110	.558	.568	.331
5	BioBERT + K means - PEFT Tuned LongT5	.181	**.028**	.134	**.591**	**.453**	.362

Table 1: Model evaluation results. BERTscore values are average F1. Values in bold indicate the best model.

a 'deep' semantic understanding of sentence meanings during clustering, and a 'broad' ability to handle extended inputs with LongT5's localized-global sparse attention.

We hypothesize that the suboptimal performance of *BioBERT + UMAP + HDBSCAN + LongT5* is due to UMAP dimensionality reduction, which might lead to the loss of critical keywords or phrases, as suggested by its lower ROUGE scores.

4.3 Improving the extractive-abstractive combination model via fine-tuning

4.3.1 Full Fine-Tuning (FFT)

After selecting *BioBERT + K-means + LongT5* as the most performant architecture, we performed full fine-tuning, FFT (Experiment 4). In all FFT cases, we observe a decrease in performance in ROUGE-1, ROUGE-2, ΔEI average, and EI macro F1 (Table 1). FFT models struggle to predict the direction of the medical outcome (EI) and cannot produce relevant tokens after abstraction (ROUGE-1, ROUGE-2). FFT's BERTscore and ROUGE-L have insignificant changes.

One hypothesis for the degradation of performance lies in the hardware-constrained input length limit. By truncating input lengths from 16,384 to 2,048 tokens during training, 71% of training examples after extractive summarization are shortened to the first 2,048 tokens for abstractive summarization (see Figure 11 in Appendix C for a visualization of the extent of truncation). If a literature review has many referenced studies, this truncation could lead to a training input-output pair where the input does not contain the signal needed to inform its output, effectively causing data quality issues. Appendix C Table 5 shows an example where an FFT-generated summary only contains information from the first study.

Additionally, in the context of FFT with truncated inputs, K-means extractive summarization has a neutral or mildly negative effect on model performance, regardless of whether it is used during training/fine-tuning, inference, or both (Appendix C, Table 3), which contradicts our non-FFT model learnings (adding K-means extractive summarization improves LongT5). One possible explanation is that FFT models can exhibit extractive capabilities by adjusting their attention weights to focus on crucial tokens or phrases needed for medical summaries, negating the need for K-means extractive summarization.

4.3.2 Parameter-Efficient Fine Tuning (PEFT)

In *BioBERT + K-means + PEFT LongT5* (Experiment 5), we see a marked improvement in BERTscore, ΔEI, and ROUGE-L compared to its non-finetuned counterpart (Table 1). These improvements highlight the importance of having a large input length and PEFT's role in enabling it.

However, the PEFT model underperformed relative to the non-finetuned variant in EI macro-averaged F1, despite an observed decrease (improvement) in the ΔEI average. ΔEI, a classification-based evaluation, classifies each target and generated summary into three evidence directions. Analysis of the classification report for summaries generated by the PEFT model (Appendix C, Table 4) revealed a notable class imbalance, dominated by targets classified as "no significant difference." This imbalance in the training dataset led to improved scores for the majority class ("no significant difference") but also deteriorated scores for the other two classes, leading to a lower macro F1. The PEFT model takes shortcuts to cross-entropy loss by over-generating summaries that imply insignificant medical outcomes. Future work should address this class imbalance using traditional class-balancing techniques (downsampling, class-balanced loss).

Oddly, after FFT or PEFT/LoRA, generated

summaries are shorter than those generated by its non-finetuned version (Experiment 1) or target summaries (Figure 2). The distribution of finetuned summary lengths centers around the targets' mode length, which means that the finetuned models learned the most common positional information for generating a stop token. Our finetuned models cannot generate more extended summaries (e.g., medical reviews that investigate multiple outcomes). Unsurprisingly, this effect increases precision but decreases recall for ROUGE scores and BERTscore (Appendix C, Figure 12).

At a high level, an example-level difference in metric scores between pairs of models is available in Appendix D.

5 Discussion

5.1 Balancing Factual Accuracy with Semantic Coherence

Our study highlights the complex relationship between factual accuracy and semantic coherence in biomedical summarization. The impressive performance of the *PEFT BioBERT + K-means + LongT5* model demonstrates the benefits of a comprehensive approach that melds context understanding, efficient data clustering, and advanced summarization techniques. Notably, the model's memory efficiency plays a critical role in processing larger data sets, significantly enhancing the quality of our summaries. This underlines the importance of accommodating substantial input lengths in MDS tasks. The findings emphasize the crucial influence of computational efficiency in improving the outputs of such models.

5.2 Comparative Analysis of Extractive and Abstractive Models

The BioBERT + K-means model maintained factual accuracy, as evidenced by its superior ΔEI score. In contrast, the LongT5 model demonstrated prowess in generating contextually rich summaries. The mixed attention mechanism of LongT5 was crucial in efficiently processing larger datasets.

5.3 Dealing with Diverse Summaries in the MS^2 Dataset

The MS^2 dataset, notable for its varied summary targets, poses significant challenges. Analysis (Appendix C Table 6) shows that the PEFT *BioBERT + K-means + LongT5* model excels with shorter summaries, as indicated by high BERTscores, but struggles with longer ones. The dataset's target sentences, derived from review article abstracts using a SciBERT-based model and human annotations, sometimes inadvertently include complex parts from the results section, leading to target length irregularity. Improving the consistency of target generation could improve data quality and increase the relevance of our summarization efforts.

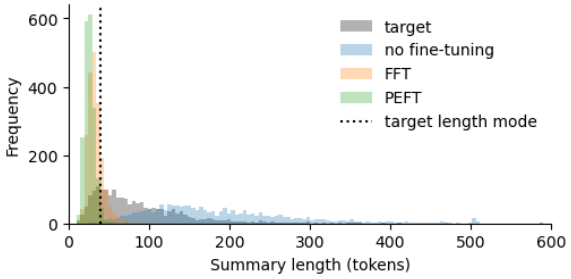

Figure 2: Comparison of test dataset summary length (tokens) between targets and variations of *BioBERT K-means + LongT5*: no fine-tuning, FFT with K-means input, and PEFT.

5.4 Implications and Future Directions

We propose exploring **Retrieval-Augmented Generation (RAG) methodologies** that leverage background information (provided in MS^2) as a query basis to extract targeted information in studies' abstracts (key and value). It is unclear if this method could rival BioBERT-K-means extraction.

Drawing inspiration from BioBART, a **specialized LongT5 architecture fine-tuned for biomedical terminology,** could help retain biomedical words through abstraction decoding. A "BioLongT5" would integrate the extensive text processing capabilities of LongT5 with a deep semantic understanding of the biomedical lexicon. Additionally, a **pointer generator** method could address the vocabulary limitations of specialized biomedical terms.

Lastly, incorporating **custom mixed loss functions with ΔEI** into the training pipeline can minimize divergence in evidence directions alongside traditional cross-entropy loss. Researchers might use LongT5's encoder hidden states for classification, offering a less resource-intensive alternative to ΔEI's RoBERTa.

6 Conclusion

In summary, our project has made strides in the application of multi-document summarization (MDS) techniques in the biomedical domain, targeting

conciseness, semantic similarity, and medical evidence directions. Our research focused on hybrid extractive (BioBERT and K-means clustering) and abstractive (LongT5) methods. We also fine-tuned our hybrid model using full fine-tuning and parameter-efficient fine-tuning (PEFT) and highlighted some challenges with fine-tuning.

The comparative analysis of our models, particularly the PEFT BioBERT + K-means + LongT5 model, has shown impressive results in balancing factual accuracy with semantic coherence. The success of this model emphasizes the critical role of large enough input lengths during fine-tuning, and the importance of memory efficiency through PEFT to enable it.

References

[1] Dong Lu and Charlin. Multi-xscience. *arXiv*, 2020.

[2] Byron C. Wallace, Sayantan Saha, Frank Soboczenski, and Iain J. Marshall. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. 2020.

[3] Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. MSˆ2: Multi-document summarization of medical studies. pages 7494–7513, November 2021.

[4] Yulia Otmakhova, Karin Verspoor, Timothy Baldwin, Antonio Jimeno Yepes, and Jey Han Lau. M3: Multi-level dataset for multi-document summarisation of medical studies. pages 3887–3901, December 2022.

[5] Rahul Tangsali, Aditya Jagdish Vyawahare, Aditya Vyankatesh Mandke, Onkar Rupesh Litake, and Dipali Dattatray Kadam. Abstractive approaches to multidocument summarization of medical literature reviews. pages 199–203, October 2022.

[6] Kartik Shinde, Trinita Roy, and Tirthankar Ghosal. An extractive-abstractive approach for multi-document summarization of scientific articles for literature review. pages 204–209, October 2022.

[7] Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. Multi-document summarization via deep learning techniques: A survey. *ACM Comput. Surv.*, 55(5), dec 2022.

[8] Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. 2018.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019.

[10] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019.

[11] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. 2019.

[12] Derek Miller. Leveraging bert for extractive text summarization on lectures. 2019.

[13] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. 2020.

[14] Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. Biobart: Pretraining and evaluation of a biomedical generative language model. 2022.

[15] Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. Longt5: Efficient text-to-text transformer for long sequences. 2022.

[16] Muhammad Sidik Asyaky and Rila Mandala. Improving the performance of hdbscan on short text clustering by using word embedding and umap. *IEEE*, 2021.

[17] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. 2021.

Appendix

A Overview of Baseline Models

Pegasus (Pre-training with Extracted Gap-sentences for Abstractive Summarization Sequence-to-sequence)

Developed by Google, Pegasus is a model tailored for summarization tasks. It utilizes a novel pre-training strategy, where it constructs 'gap sentences' and learns to generate these absent segments. This approach specifically enhances its proficiency in producing coherent and extensive summaries. Such a feature is particularly advantageous for summarizing complex and lengthy texts, as exemplified in the MS2 dataset.

BioBART (Biomedical Bidirectional and Auto-Regressive Transformers)

An adaptation of the BART architecture for biomedical applications, BioBART incorporates both bidirectional and autoregressive transformers. This configuration facilitates a profound comprehension

of intricate biomedical contexts. Trained extensively on biomedical literature, BioBART shows a marked aptitude for summarizing complex medical texts, aligning well with the preprocessing steps designed for the MS2 dataset, particularly in processing nuanced information within the dataset's abstracts.

T5 (Text-to-Text Transfer Transformer)

T5 is predicated on the concept that all text-based tasks are convertible into a text-to-text format. This model's versatility, coupled with its comprehensive training across various language tasks, renders it an effective tool for summarization. In biomedical summarization, T5's adaptability and linguistic sophistication are essential for generating accurate and comprehensive summaries from the MS2 dataset's diverse and complex texts.

BERT (Bidirectional Encoder Representations from Transformers)

A pioneering model in NLP, BERT was developed by Google and introduces a bidirectional training approach. Distinct from traditional unidirectional models, BERT processes text in both directions (left-to-right and right-to-left), allowing for a nuanced understanding of context and inter-sentence relationships. This characteristic makes BERT particularly effective in extractive summarization tasks, such as identifying and extracting key information from the extensive biomedical literature in the MS2 dataset.

BioBERT (Biomedical Bidirectional Encoder Representations from Transformers)

An extension of BERT, BioBERT is pre-trained on general language corpora and extensive biomedical literature. This specialized training equips BioBERT with a heightened ability to process biomedical terminology and concepts accurately. It excels in identifying and extracting vital biomedical information, making it highly suitable for summarizing scientific articles and studies within the MS2 dataset. Its domain-specific focus enhances accuracy in handling the complex details and technical language prevalent in biomedical texts.

SciBERT (Scientific Bidirectional Encoder Representations from Transformers)

SciBERT adapts the BERT model for the scientific domain. Pre-trained on a vast corpus of scientific literature, including papers and journals across various scientific fields, SciBERT acquires a deep understanding of scientific terminology and concepts. This pre-training makes it particularly adept at extractive summarization within scientific literature, enabling it to discern and extract relevant information from complex scientific texts, a capability that is crucial for summarizing the detailed and nuanced content in the MS2 dataset.

B Schematics of Baseline and Experimented Models

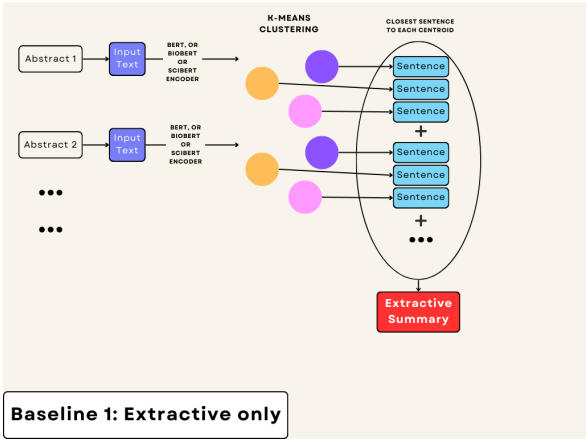

Figure 3: This diagram illustrates the use of BERT or similar models for encoding input text, followed by a K-means clustering algorithm to identify the most representative sentences, which are then used to create an extractive summary.

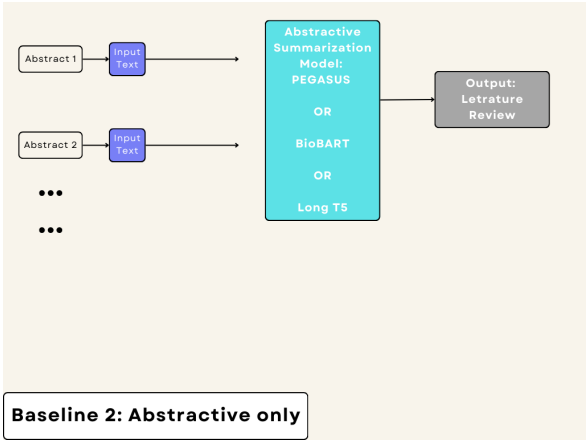

Figure 4: This diagram shows the flow of input text being processed directly by abstractive summarization models like PEGASUS, BioBART, or Long T5 to generate a literature review.

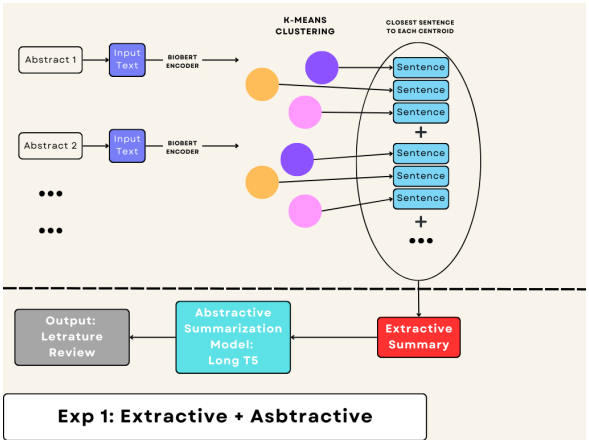

Figure 5: The third diagram represents a hybrid approach that uses BioBERT for text encoding and K-means for clustering to generate an extractive summary, which is then fed into the Long T5 model to create a literature review.

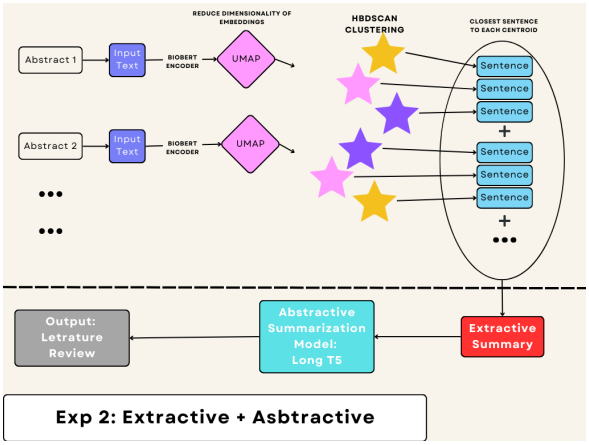

Figure 6: This diagram depicts a similar hybrid approach as Experiment 1, but with an added step of dimensionality reduction using UMAP before clustering with HDBSCAN, leading to an extractive summary that is processed by the Long T5 model.

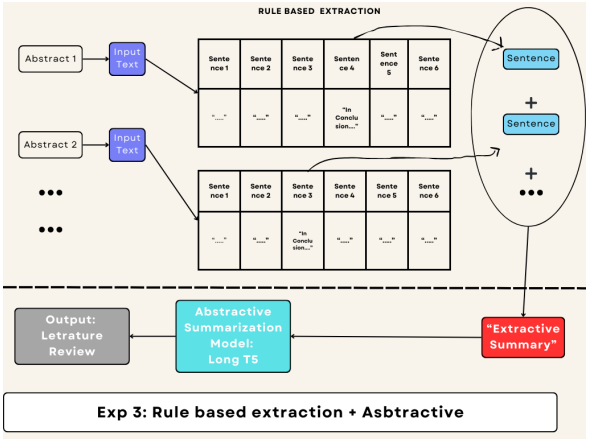

Figure 7: The fifth diagram presents another hybrid approach where rule-based sentence extraction is combined with an abstractive summarization using the Long T5 model to produce a literature review.

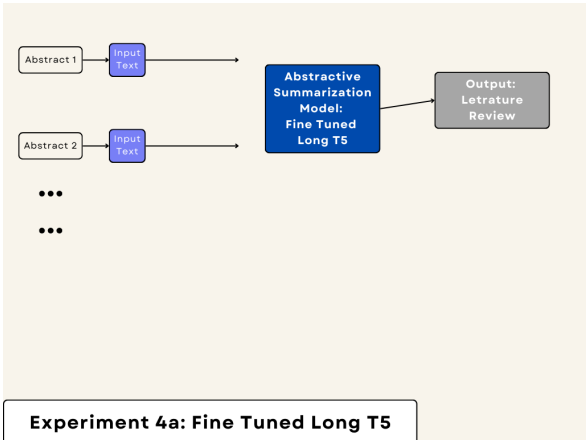

Figure 8: Here, the input text is summarized abstractively using a fine-tuned Long T5 model to produce a literature review.

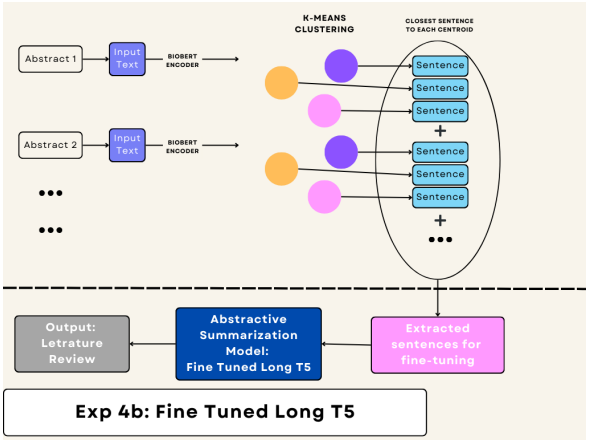

Figure 9: This diagram shows an extractive preprocessing step using BioBERT and K-means clustering to extract sentences for fine-tuning the Long T5 model, which then creates a literature review.

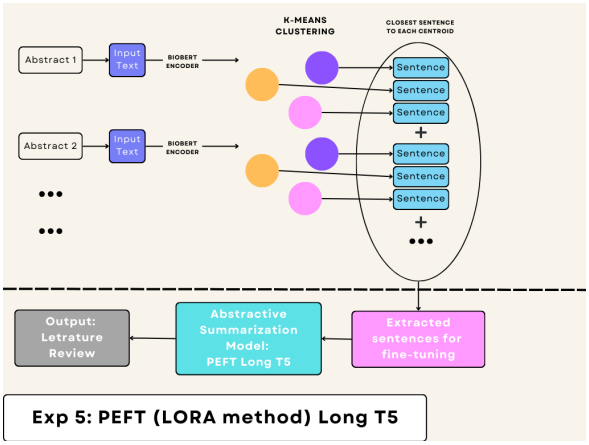

Figure 10: The final diagram describes using a BioBERT encoder with K-means clustering to perform extractive summarization. The extracted sentences are then used for fine-tuning an abstractive summarization model (PEFT Long T5) using the LORA method.

C Relevant Data

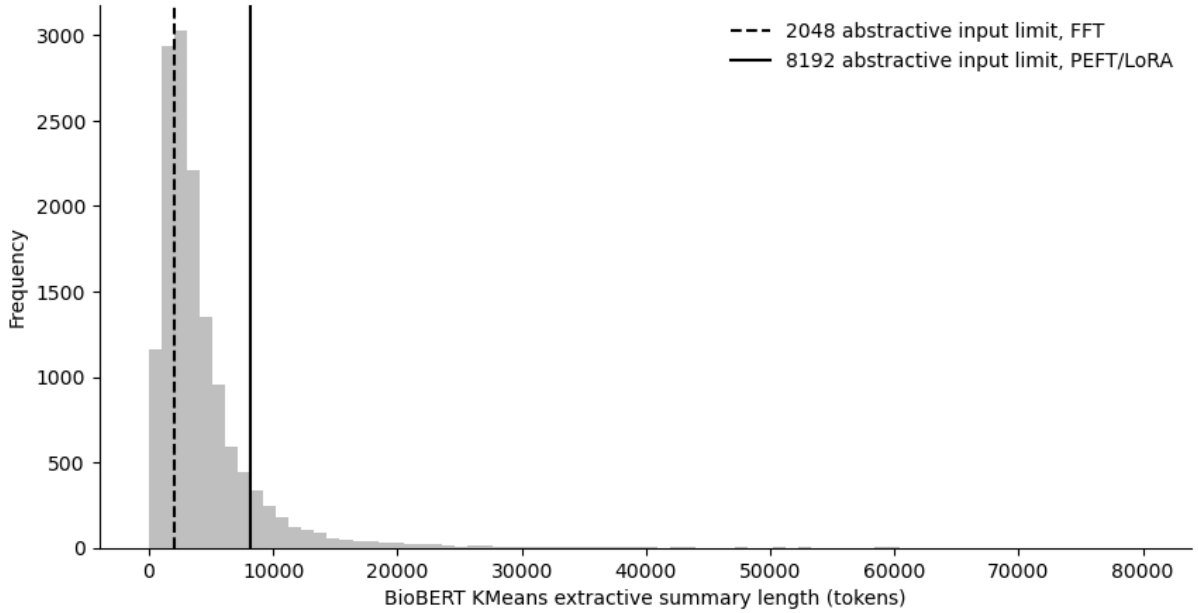

Figure 11: Histogram of lengths of BioBERT K-means extractive summaries. Vertical lines indicate input limits to abstractive summarization (2,048 for full fine-tuning, 8,192 for PEFT LoRA).

Model	R-1	R-2	R-L	B	ΔEI	F1
Pegasus (abs)	.174	.019	.109	.522	.558	.360
BioBART (abs)	.158	.036	.093	.525	.526	.378
LongT5 (abs)	.185	.022	.106	.543	.543	.364
BERT + K means (ext)	.047	.012	.030	.505	.518	.372
SciBERT + K means (ext)	.047	.012	.030	.504	.516	.376
BioBERT + K means (ext)	.046	.012	.029	.502	.503	.389

Table 2: baseline extractive and abstract Note: B = BERTscore average F1

	R-1		R-2		R-L		BS F1		ΔEI		EI F1	
	F	K	F	K	F	K	F	K	F	K	F	K
F	.154	.148	.022	.018	.115	.110	.561	.558	.576	.568	.332	.331
K	.139	.142	.018	.017	.106	.107	.560	.558	.594	.580	.321	.331

Table 3: Metrics comparison of 4 combinations of full fine-tuned (FFT) models, where either the full dataset (F) or a K-means extracted dataset (K) is used during training (rows) and inference (columns).

	BioBERT + K means + LongT5			BioBERT + K means + PEFT LongT5			
	P	R	F1	P	R	F1	Support
significantly decreased	.117	.194	.146	.103	.072	.085	111
no significant difference	.636	.548	.589	.626	.778	.694	1214
significantly increased	.386	.438	.41	.401	.25	.308	696
Accuracy			.491			.558	2021
Macro Average	.380	.393	.382	.377	.367	.362	2021
Weighted Average	.521	.491	.503	.520	.558	.528	2021

Table 4: Classification report of Evidence Inference for BioBERT K-means + LongT5, with and without PEFT (Experiments 1 and 5), based on test dataset.

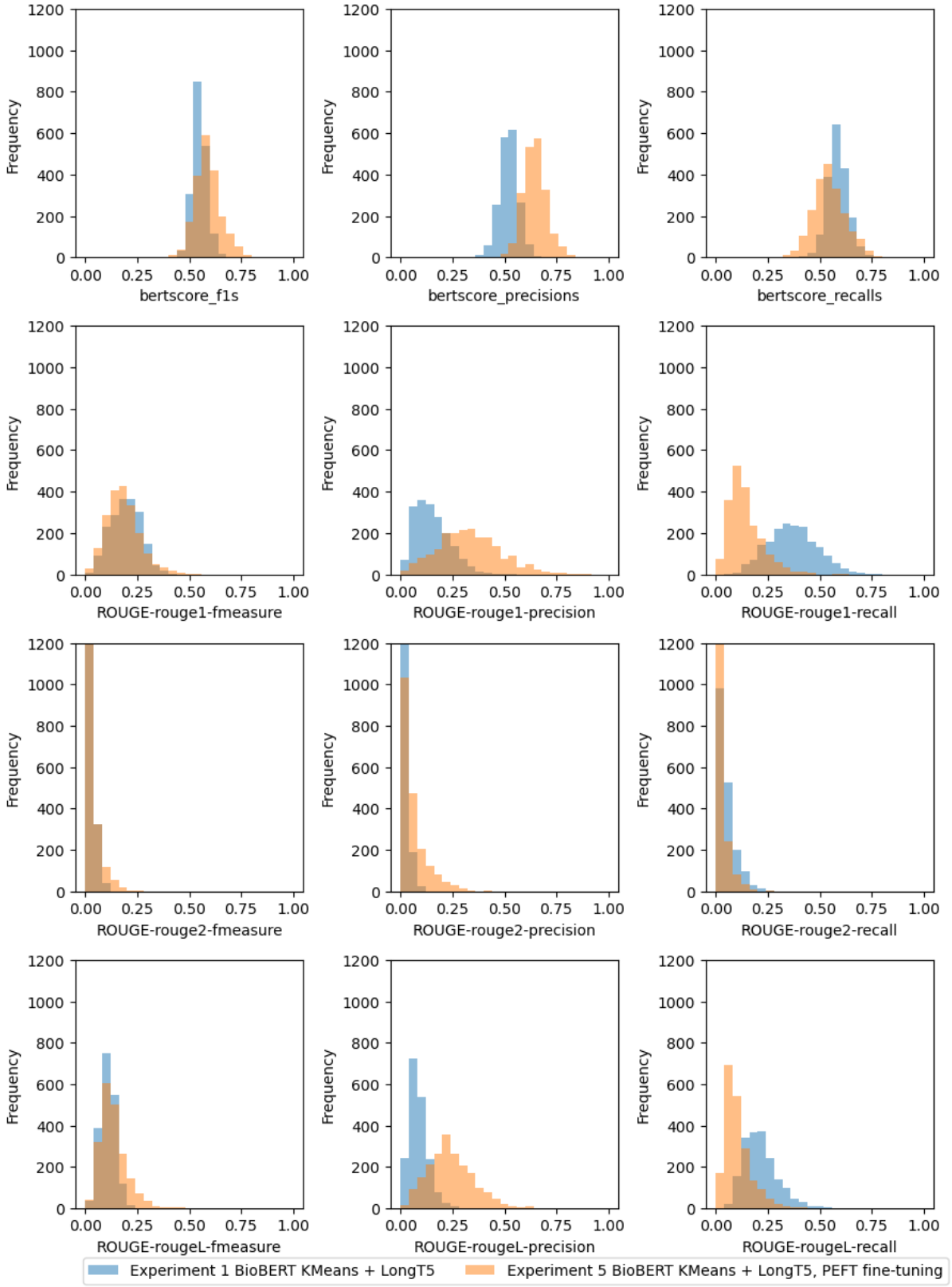

Figure 12: Histograms of BERTscore, ROUGE-1, ROUGE-2, ROUGE-L, in F1, precision and recall, for two models: Experiment 1 BioBERT K-means + LongT5 (no finetuning) and Experiment 5 BioBERT K-means + LongT5, PEFT fine-tuning. Fine-tuning with PEFT minimally affects F1, but drastically increases precision and decreases recall of generated summaries.

			Experiment 1 BioBERT KMeans + LongT5		Experiment 4 BioBERT KMeans + LongT5, FFT, KMeans input		
Review ID	Target	Abstract titles (in order of appearance)	Generated summary	ROUGE-1 F1 Score	Generated summary	ROUGE-1 F1 Score	Difference in ROUGE-1 F1
26655787	Results reveal no significant difference between the groups , with respect to BMI , while PA and DIET yielded a greater reduction in HbA1c . Significant reduction in both systolic and diastolic pressures in the DIET group , and diastolic pressure in the PA group , was observed . HDL-c in the DIET group was significantly higher than the control group , while no change in LDL-c levels , was seen in all three intervention subtypes . There was no difference between the EDU vs. the control group in terms of HbA1c , blood pressure or HDL-c and LDL-c . CONCLUSION DIET intervention showed an improvement in HbA1c , systolic/diastolic blood pressure and HDL-c , with an exception of LDL-c and BMI , suggesting that nutritional intervention had a significant impact on the quality of life by reducing the cardiovascular risk in type 2 diabetes patients	['Effects of a structured health education programme by a diabetic education nurse on cardiovascular risk factors in Chinese Type 2 diabetic patients: a 1-year prospective randomized study.', 'Weight Loss, Glycemic Control, and Cardiovascular Disease Risk Factors in Response to Differential Diet Composition in a Weight Loss Program in Type 2 Diabetes: A Randomized Controlled Trial', 'Effect of Exercise on Blood Pressure in Type 2 Diabetes: A Randomized Controlled Trial', 'Effects of a 12-month physical activity counselling intervention on glycaemic control and on the status of cardiovascular risk factors in people with Type 2 diabetes', 'Biophysiologic outcomes of the Enhancing Adherence in Type 2 Diabetes (ENHANCE) trial.', 'Long-term lifestyle intervention lowers the incidence of stroke in Japanese patients with type 2 diabetes: a nationwide multicentre randomised controlled trial (the Japan Diabetes Complications Study)', 'Case management for patients with poorly controlled diabetes: a randomized trial.', 'A Pharmacist Care Program', 'The Cochrane Collaboration's tool for assessing risk of bias in randomised trials', 'Reduction in Weight and Cardiovascular Disease Risk Factors in Individuals With Type 2 Diabetes', 'Effectiveness of PRECEDE model for health education on changes and level of control of HbA1c, blood pressure, lipids, and body mass index in patients with type 2 diabetes mellitus', 'Long-term effects of a lifestyle intervention on weight and cardiovascular risk factors in individuals with type 2 diabetes mellitus: four-year results of the Look AHEAD trial.', 'Prescription of physical activity is not sufficient to change sedentary behavior and improve glycemic control in type 2 diabetes patients.', 'Multifactorial intervention in individuals with type 2 diabetes and microalbuminuria: the Microalbuminuria Education and Medication Optimisation (MEMO) study.', 'Lifestyle intervention by group care prevents deterioration of Type II diabetes: a 4-year randomized controlled clinical trial', 'The anti-inflammatory effects of exercise training in patients with type 2 diabetes mellitus', 'Effect of an intensive exercise intervention strategy on modifiable cardiovascular risk factors in subjects with type 2 diabetes mellitus: a randomized controlled trial: the Italian Diabetes and Exercise Study (IDES).', 'Nutritional intervention in patients with type 2 diabetes who are hyperglycaemic despite optimised drug treatment—Lifestyle Over and Above Drugs in Diabetes (LOADD) study: randomised controlled trial', 'Culturally sensitive patient-centred educational programme for self-management of type 2 diabetes: a randomized controlled trial.']	The effect of regular exercise on blood pressure, body mass, and lipids in patients with Type 2 Diabetes is investigated. In the intervention group, physical activity was shown to be an effective treatment for both glyceric control and cardiovascular risk factors. However, there was no significant change in other measures of control during the trial period. Furthermore, stroke was significantly reduced in the pre-intervention group compared to that of the control group. Precede health education model was a useful tool in improving metabolic control as well as the reduction of heart disease in type 2 patients.	0.415	Conclusions : Chinese herbal education has the potential to reduce blood pressure in DDM patients.	0.078	-0.337

Table 5: Example of a summary where full fine-tuning produces a worse ROUGE-1 score than the non-fine-tuned model (BioBERT K-means + LongT5). The full fine-tuned (FFT) model shows a tendency to include information from earlier studies. In this example, the FFT summary centers on “Chinese” and “education,”, which are topics related to the first study.

Review ID	Background	Target	Experiment 5 BioBERT KMeans + LongT5, PEFT summary	BERTscore F1
20927745	BACKGROUND Benign prostatic hyperplasia (BPH) , a non-malignant enlargement of the prostate in aging men , can cause bothersome urinary symptoms (intermittency , weak stream , straining , urgency , frequency , incomplete emptying) . Finasteride , a five-alpha reductase inhibitor (5ARI) , blocks the conversion of testosterone to dihydrotestosterone , reduces prostate size , and is commonly used to treat symptoms associated with BPH . OBJECTIVES To compare the clinical effectiveness and harms of finasteride versus placebo and active controls in the treatment of lower urinary tract symptoms (LUTS) .	MAIN RESULTS Finasteride consistently improved urinary symptom scores more than placebo in trials of > 1 year duration , and significantly lowered the risk of BPH progression (acute urinary retention , risk of surgical intervention , ≥ 4 point increase in the AUASI/IPSS) . In comparison to alpha-blocker monotherapy , finasteride was less effective than either doxazosin or terazosin , but equally effective compared to tamsulosin . Both doxazosin and terazosin were significantly more likely than finasteride to improve peak urine flow and nocturia , versus finasteride . Versus tamsulosin , peak urine flow and QoL improved equally well versus finasteride . However , finasteride was associated with a lower risk of surgical intervention compared to doxazosin , but not to terazosin , while finasteride and doxazosin were no different for risk of acute urinary retention . Finasteride + doxazosin and doxazosin monotherapy improved urinary symptoms equally well (≥ 4 point improvement).For finasteride , there was an increased risk of ejaculation disorder , impotence , and lowered libido , versus placebo . Versus doxazosin , finasteride had a lower risk of asthenia , dizziness , and postural hypotension , and versus terazosin , finasteride had a significant , lower risk of asthenia , dizziness , and postural hypotension . Finasteride improves long-term urinary symptoms versus placebo , but is less effective than doxazosin . Long-term combination therapy with alpha blockers (doxazosin , terazosin) improves symptoms significantly better than finasteride monotherapy . Finasteride + doxazosin improves symptoms equally - and clinical ly - to doxazosin alone . Finasteride + doxazosin versus doxazosin improves scores equally for short and long term . Drug-related adverse effects for finasteride are rare ; nevertheless , men taking finasteride are at increased risk for impotence , erectile dysfunction , decreased libido , and ejaculation disorder , versus placebo . Versus doxazosin , which has higher rates of dizziness , postural hypotension , and asthenia , men taking finasteride are at increased risk for impotence , erectile dysfunction , decreased libido , and ejaculation disorder . Finasteride significantly reduces asthenia , postural hypotension , and dizziness versus terazosin . Finasteride significantly lowers the risk of asthenia , dizziness , ejaculation disorder , and postural hypotension , versus finasteride + terazosin	There was no statistically significant difference in the risk of sexual adverse events.	0.3786
22002191	OBJECTIVE To conduct a systematic review to evaluate the evidence of the use of incentive spirometry (IS) for the prevention of postoperative pulmonary complications and for the recovery of pulmonary function in patients undergoing abdominal , cardiac and thoracic surgeries .	CONCLUSION There was no evidence to support the use of incentive spirometry in the management of surgical patients .	Conclusions : There is no evidence to support the use of inspiratory breath exercises as an alternative to physical therapy after abdominal surgery	0.8246

Table 6: Two examples from the test dataset showing worst and best BERTscore F1 for BioBERT K-means + LongT5, PEFT fine-tuning (Experiment 5)

D Example-level metric comparison between pairs of models

For each subsequent plot, we present scatterplots of scores (ROUGE-1, ROUGE-2, ROUGE-L, BERTscore, ΔEI) between pairs of models in off-diagonal subplots. Each scatterplot also shows an x=y parity line (examples on this line have the same score in both models), and a red "X" indicating the mean of scores. Diagonal subplots contain histograms of scores.

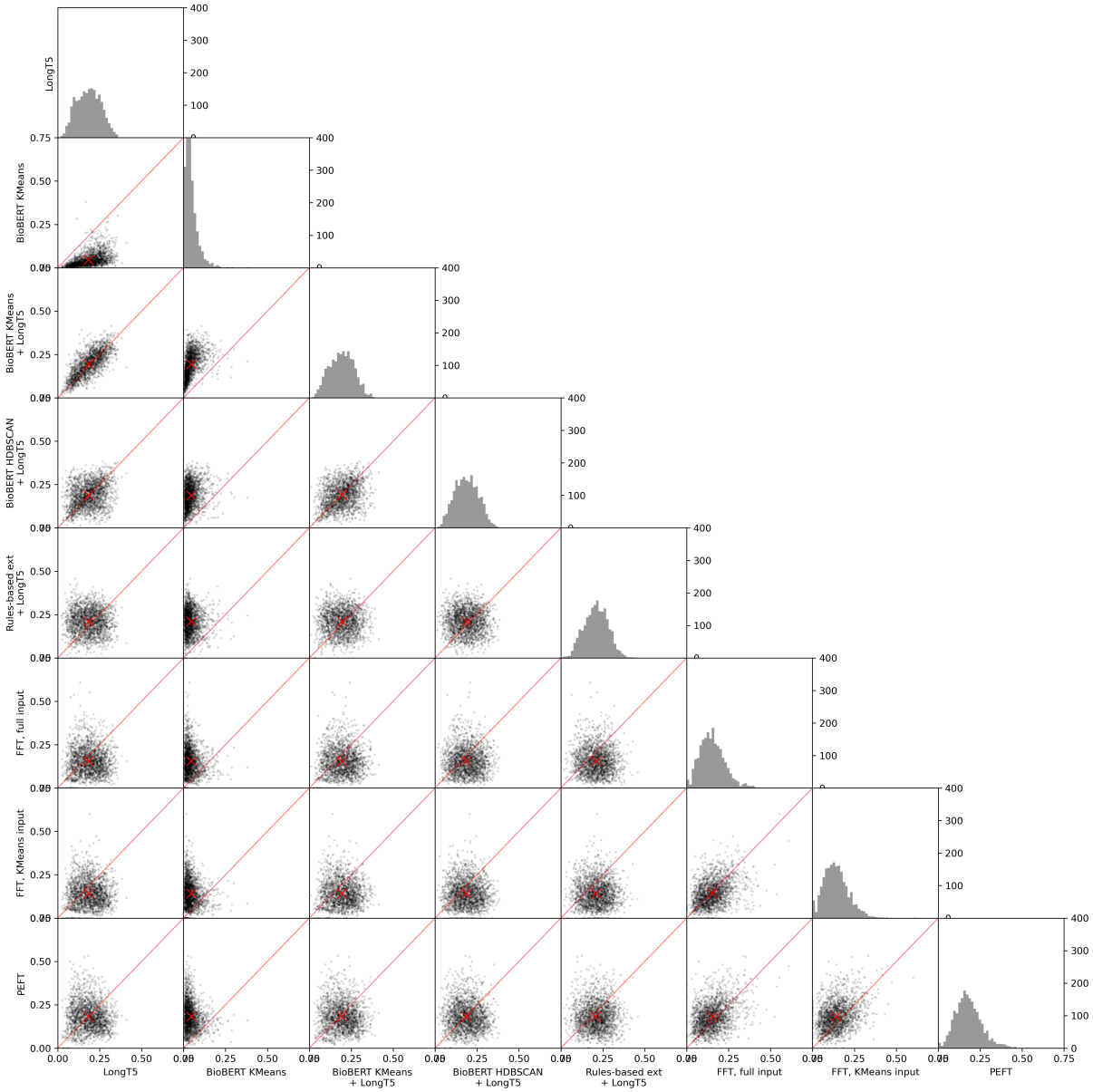

Figure 13: Example-level comparison of ROUGE-1 F1.

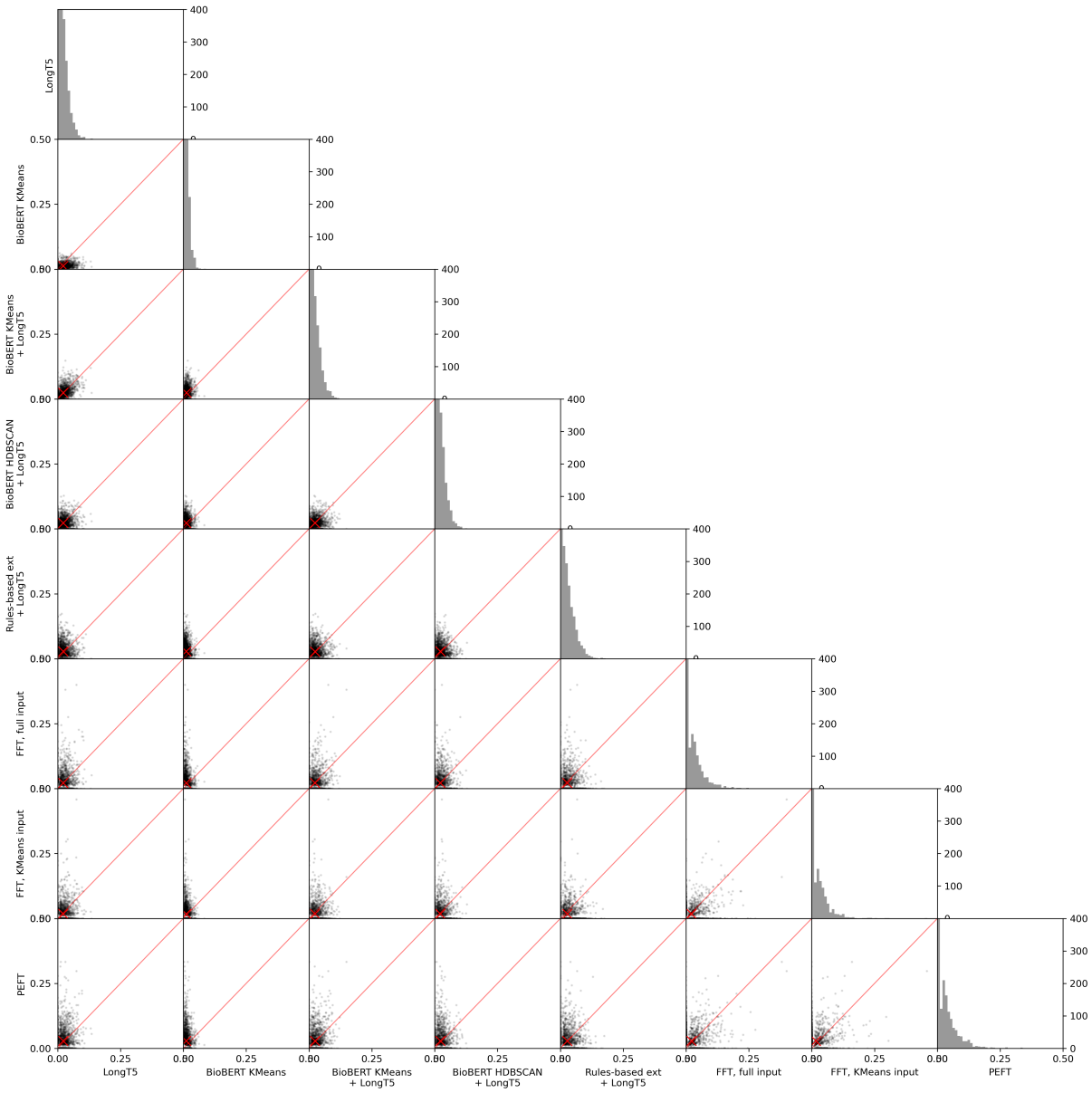

Figure 14: Example-level comparison of ROUGE-2 F1.

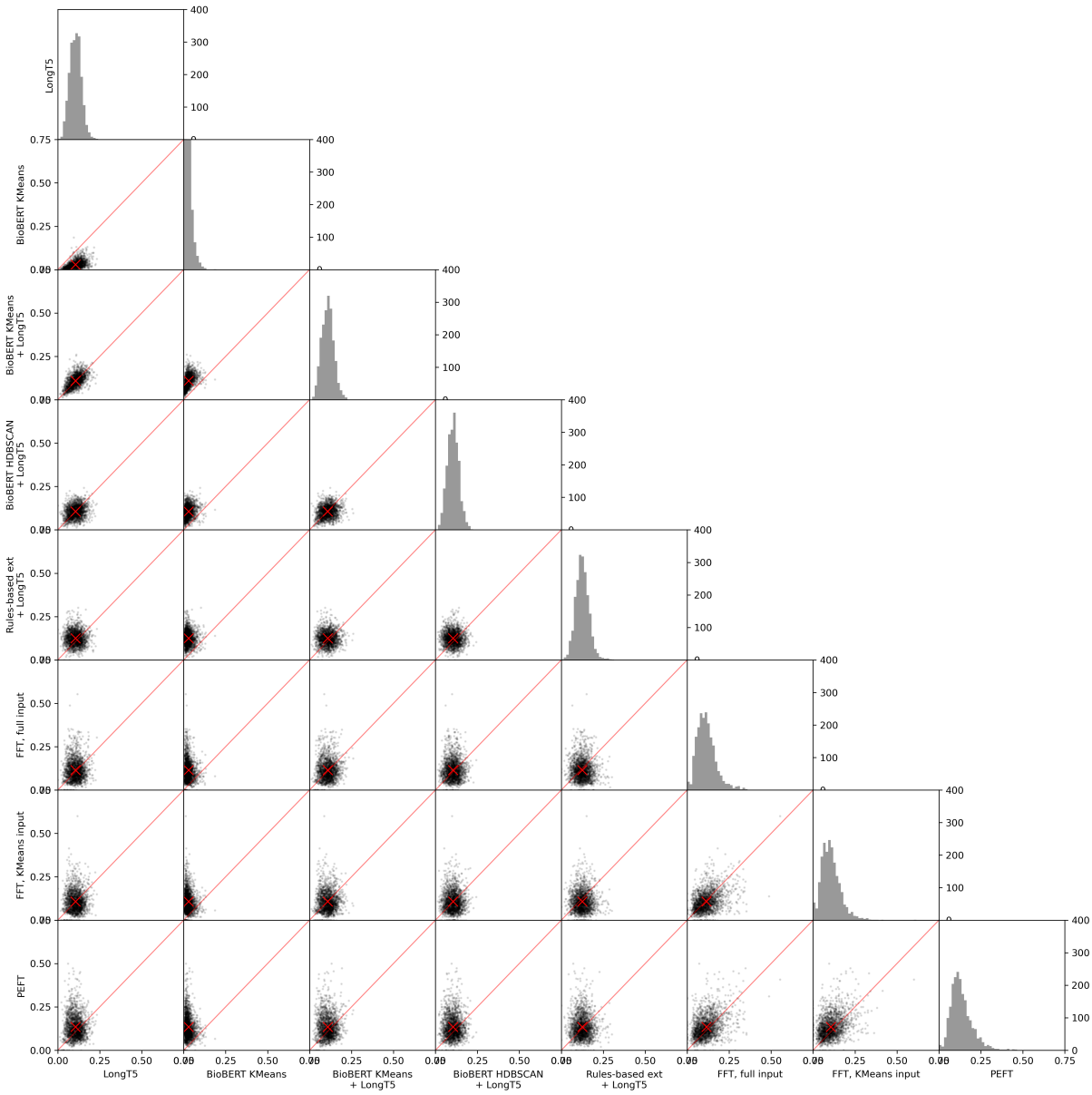

Figure 15: Example-level comparison of ROUGE-L F1.

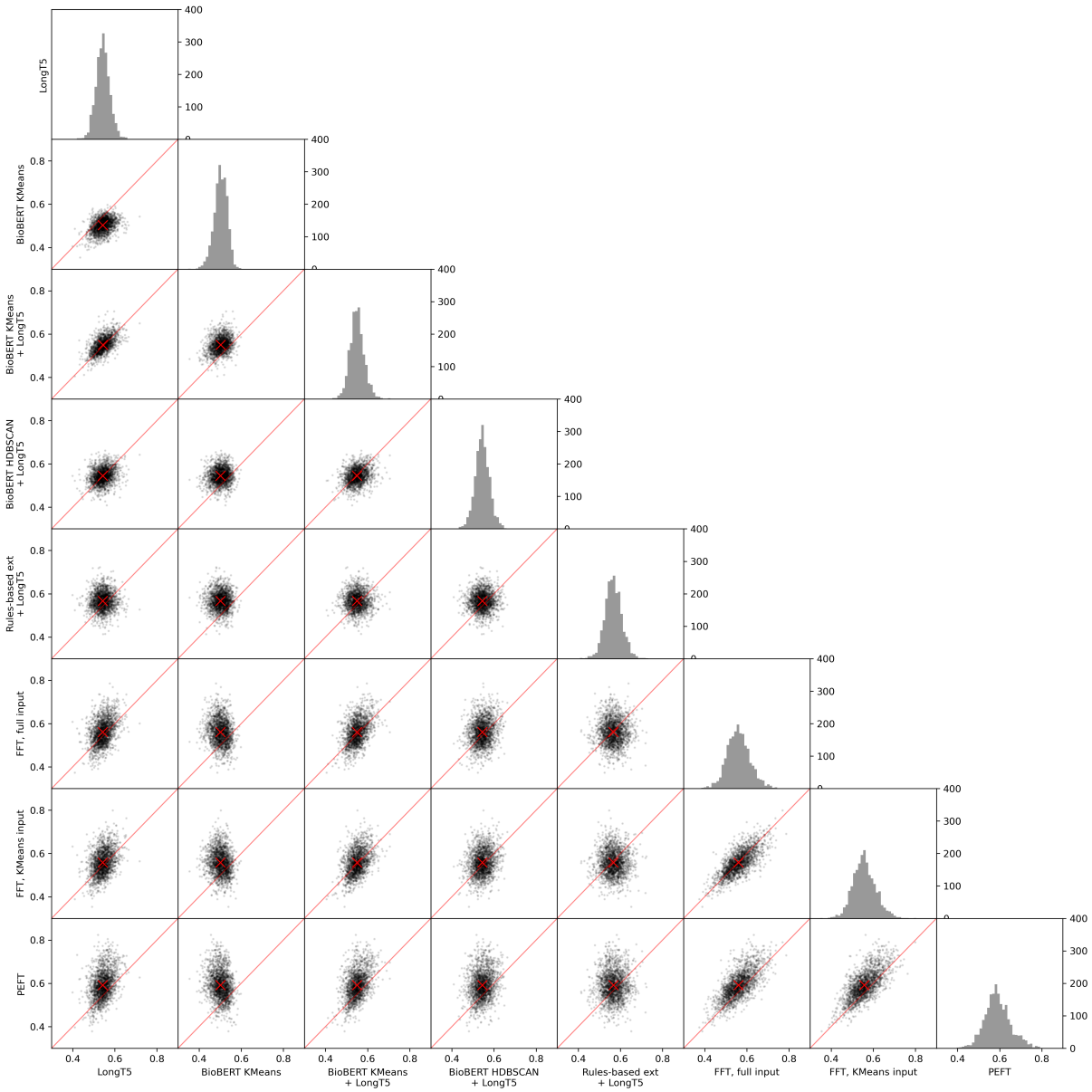

Figure 16: Example-level comparison of BERTscore F1.

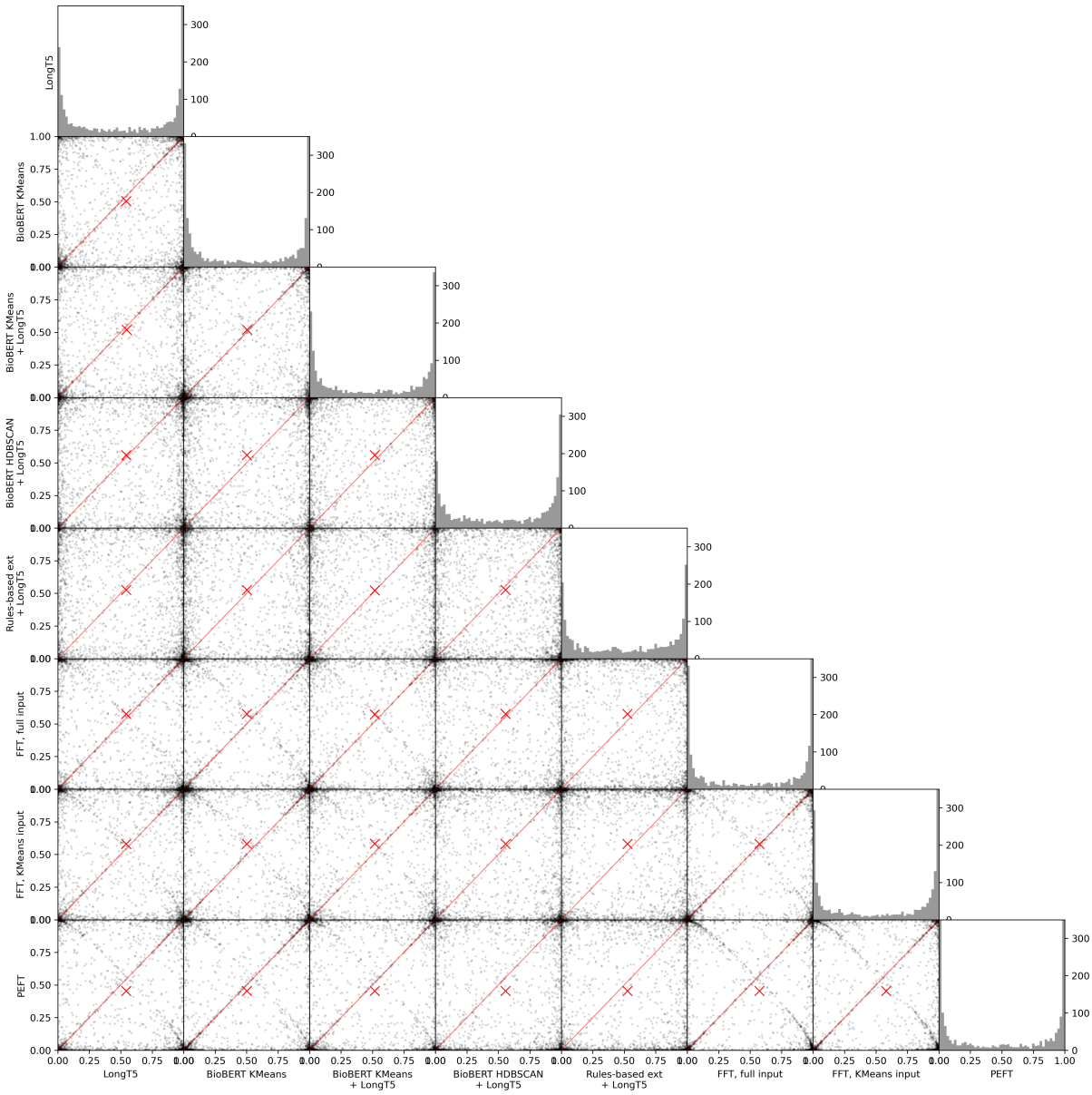

Figure 17: Example-level comparison of ΔEI.