BioLay_AK_SS at BioLaySumm: Domain Adaptation by Two-Stage Fine-Tuning of Large Language Models used for Biomedical Lay Summary Generation

Akanksha Karotia, Seba Susan
Department of Information Technology, Delhi Technological University, Delhi, India
akankshakarotia@gmail.com, seba_406@yahoo.in

Abstract

Lay summarization is essential but challenging, as it simplifies scientific information for non-experts and keeps them updated with the latest scientific knowledge. In our participation in the Shared Task: Lay Summarization of Biomedical Research Articles @ BioNLP Workshop (Goldsack et al., 2024), ACL 2024, we conducted a comprehensive evaluation on abstractive summarization of biomedical literature using Large Language Models (LLMs) and assessed the performance using ten metrics across three categories: relevance, readability, and factuality, using eLife and PLOS datasets provided by the organizers. We developed a two-stage framework for lay summarization of biomedical scientific articles. In the first stage, we generated summaries using BART and PEGASUS LLMs by fine-tuning them on the given datasets. In the second stage, we combined the generated summaries and input them to BioBART, and then fine-tuned it on the same datasets. Our findings show that combining general and domain-specific LLMs enhances performance.

1 Introduction

In today's era, a lot of research is being conducted in the field of biomedical science, resulting in a huge amount of biomedical literature. The vast scientific knowledge poses a challenge for healthcare professionals, researchers, and the non-expert public in staying informed about advancements in the biomedical domain (Bishop et al., 2022; Karotia and Susan, 2023). Making the information accessible and understandable, regardless of their background knowledge, is difficult. Manually summarizing long scientific articles requires too much domain-oriented knowledge, effort, and time, especially for lay summarization. First, summarizing and then transforming the summarized information for non-experts is impractical. This problem can be tackled by designing lay summarization systems that bridge the gap between non-experts and experts by modifying intricate scientific knowledge into a clear and condensed form with increased readability. This step will increase scientific literacy and enable decision-making for experts and non-experts.

This study's contributions include:

- In the first phase of the model, BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020) general domain LLMs were used. These LLMs were fine-tuned on eLife and PLOS datasets for the summarization task.
- The outputs from both LLMs are combined, and sentences are deduplicated to eliminate redundant data and enhance diversity and inclusivity.
- In the second phase, the deduplicated data was sent to the BioBART (Yuan et al., 2022) LLM, which is pre-trained on biomedical datasets and further fine-tuned by the authors on the dataset made accessible by the challenge organizers.
- Performance evaluation and analyses are done for relevance, readability, and factuality metrics.

2 Related Work

Recent studies have showcased the significant potential of large language models (LLMs) in natural language generation tasks. In addition, the first version of the BioLaySum (Goldsack et al., 2023) illustrated the effectiveness of utilizing LLMs in both summary formation (Turbitt et al., 2023) and data augmentation (Sim et al., 2023). LLMs exhibit proficiency in perceiving complex relationship patterns due to their training on diverse large-scale datasets across various tasks (Karotia and Susan, 2022).

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 762–768
August 16, 2024. ©2024 Association for Computational Linguistics

(Liu et al., 2023) utilized two different LLMs, BART and PEGASUS, for the BioLaySum task at ACL 2023, respectively focusing on the eLife and PLOS datasets, aiming to optimize memory usage. (Phan et al., 2023) processed long documents using an effective framework comprising of BioBART and a factorized energy-centric method. (Turbitt et al., 2023) performed comprehensive experiments with general and domain-specific GPT models using zero-shot and few-shot methods. (Al-Hussaini et al., 2023) utilized T5 and BART LLMs with an attention mechanism for information selection and further applied a zero-shot method for language simplification. (Reddy et al., 2023) employed BART to generate summaries by incorporating sentence labels, which significantly improved results. (Chen et al., 2023) used various models for each submission, including PRIMERA, PEGASUS, and BART-LongFormer.

3 Methodology

In recent times, Large Language Models such as BioBART pre-trained on biomedical corpora have achieved enhanced performance for biomedical natural language generation tasks. However, the readability of the generated summaries needs to be improved from the perspective of a non-expert audience. To achieve this aim, a two-stage framework is designed in this work, as shown in Figure 1, to generate lay summaries for complex and lengthy scientific research articles, targeting non-expert audiences. In the first phase of the framework, BART and PEGASUS general-purpose LLMs are selected for summary generation by fine-tuning them on the challenge datasets: eLife and PLOS. Both models performed well on validation and test sets, indicating their capability to generate high-quality summaries due to their pre-training on multiple tasks and large datasets. As these transformers have limited input lengths, the first 1024 tokens are used for training because these LLMs are resource-intensive and time-consuming. Specifically, training with starting information proves to be more efficient, as studies indicate that important information is typically presented at the beginning and end of research articles (Cai et al., 2022). BART and PEGASUS are transformer-based models that excel in text generation and are specifically designed for abstractive summarization. In the initial phase of our approach, we fine-tune these models using article-lay summary pairs from the eLife and PLOS datasets separately. This process involves setting the hyperparameters specified in Table 1, as discussed in Section 4. After fine-tuning, the summaries generated by both models are merged. The aggregated text is further processed to handle redundant information, which is eliminated through sentence deduplication. This results in diverse and non-redundant data, making it suitable to be input into the second phase of the framework for further fine-tuning.

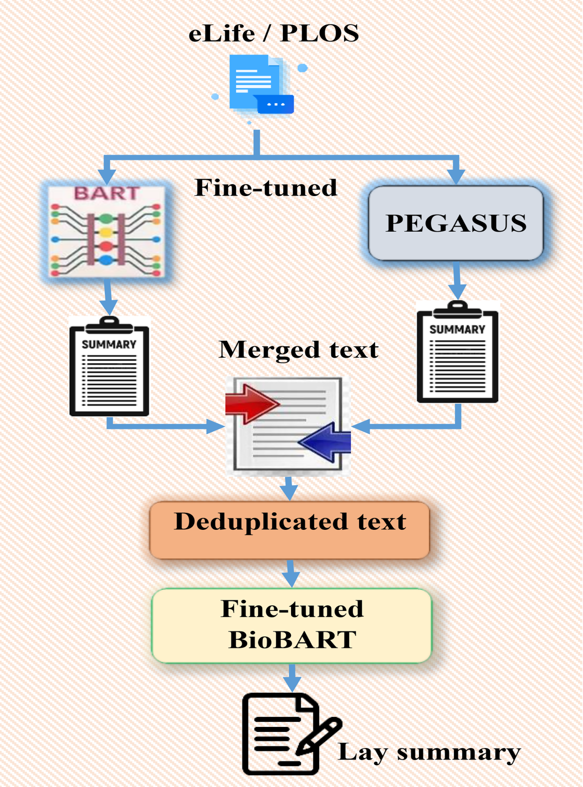

Figure 1: Proposed framework.

The authors observed two cases of redundancy after summary generation in the first phase, leading to the need for deduplication of sentences to ensure non-redundant information for the second phase. First, identical sentences are present within the same summary generated by the models. Second, the summaries generated by both models have identical sentences. This issue does not arise for all samples. But even in a few cases, it is important to ensure that non-redundant information is used for further processing to generate a quality summary.

Algorithm 1 outlines the steps for deduplicating sentences in aggregated text. This process removes identical sentences while considering case sensitivity (lower-case), ensuring non-redundant information. In the second and final phase of the framework, the deduplicated text for a correspond-

Algorithm 1: Sentence Deduplication

```
Data: $C_{txt}$ (Aggregated text)
Result: $D_{txt}$ (Deduplicated text)
$S_{tokenized} \leftarrow$ sentence_tokenize($C_{txt}$) ;
$N_{non\_duplicate\_sentences} \leftarrow$ empty list to
    store non-duplicate sentences;
$Duplicate_{ssentences} \leftarrow$ empty list to store
    duplicate sentences;
for each s in $S_{tokenized}$ do
    if s is not in $N_{non\_duplicate\_sentences}$
      then
        Append s to
          $N_{non\_duplicate\_sentences}$;
    else
        Append s to $Duplicates_{sentences}$;
    end
end
$D_{txt} \leftarrow$
    Concatenate($N_{non\_duplicate\_sentences}$);
$D_{txt}$
```

ing document is considered salient information that guides the domain-specific LLM to generate more accurate summaries. For this phase, the BioBART LLM (Yuan et al., 2022) is selected for fine-tuning, as this model is pre-trained on a vast amount of biomedical datasets and showed promising results in the first edition of the BioLaySum task. In place of the original article, the deduplicated lay summary generated in the first phase is used as input for fine-tuning BIOBART in the second phase of the proposed model. The summary obtained after fine-tuning with BioBART results in improved performance on both datasets.

Hyperparameter	**Values**
Batch size	16
Learning rate	5e-5
Early stopping	3
Max input length	1024
Min and max target length (eLife)	350, 512
Min and max target length (PLOS)	180, 200
No. of beams	4
Penalty	2

Table 1: Hyperparameter settings for all the baseline methods and the proposed model.

4 Experimental Setup

The experiments in this study are conducted on the Google Colab Pro Plus platform, using an NVIDIA A100 GPU with 40 GB of GPU RAM for training and inferencing. Appendix A provides insights into the datasets used, while Table 1 details the hyperparameter settings for all the models.

4.1 Datasets

The organizers provided the two biomedical datasets, eLife (Goldsack et al., 2022) and The Public Library of Science (PLOS) (Goldsack et al., 2022), used in this study, across which all the models were trained and evaluated. Appendix A shows the detailed dataset statistics.

4.2 Baseline and Hyperparameter Settings

[1]PEGASUS (Zhang et al., 2020): This model is pre-trained on the C4 and HugeNews datasets for abstractive summarization by utilizing the gap sentence ratio methodology, and stochastic sampling was employed for key sentence identification. It is fine-tuned on eLife and PLOS with 10 and 8 epochs, respectively.

[2]BART (Lewis et al., 2020): This model was pre-trained for language generation and translation tasks in English, and fine-tuned on the CNN/DM dataset, specifically for summarization purposes. It is fine-tuned for 13 epochs on eLife and 12 epochs on PLOS.

[3]T5-small (Raffel et al., 2020): This model, a smaller version of T5, is pre-trained on the C4 dataset for varied tasks that include paraphrasing and natural language generation. It is fine-tuned with 15 epochs on eLife and 11 epochs on the PLOS dataset.

[4]BIOBART (Yuan et al., 2022): BioBART has efficiently adapted the BART framework for generative tasks specifically tailored to the biomedical domain. The model is pre-trained on several language generation tasks, including: 1) A medical dialogue system, where the objective is to emulate a human doctor communicating with real patients, trained using the CovidDialog dataset. 2) Abstractive summarization on the iCliniq, HealthCareMagic, and MeQSum datasets. 3) Entity linking on the MedMentions, BC5CDR, and AskAPatients datasets. 4) Named entity recognition on

[1] https://huggingface.co/google/pegasus-cnn_dailymail
[2] https://huggingface.co/facebook/bart-large-cnn
[3] https://huggingface.co/google-t5/t5-small
[4] https://huggingface.co/GanjinZero/biobart-large

the ShARe13 and CADEC datasets. This model is fine-tuned with 7 epochs on eLife and 5 epochs on the PLOS dataset. The proposed model is also fine-tuned for 12 epochs using the same parameters listed in Table 1.

4.3 Evaluation metrics

Various metrics have been employed to evaluate the model's performance, categorized into three main aspects: relevance, readability, and factuality.

Relevance: Four metrics are employed to assess the relevance aspect: ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004), and BERTScore (Zhang et al.). Higher scores on these metrics indicate better performance.

Readability: The Flesch-Kincaid Grade Level (FKGL), Dale-Chall Readability Score (DCRS), Coleman-Liau Index (CLI), and LENS are used as readability metrics. Lower scores for FKGL, DCRS, and CLI indicate better readability, whereas higher scores for LENS are considered.

Factuality: AlignScore and SummaC are the metrics evaluated to measure the factual accuracy of the generated summaries. Higher scores on these metrics indicate higher quality in terms of factuality.

4.4 Results and Discussion

The performance of the baseline methods and the system proposed in this study is shown in Table 2 and Table 3 for the validation and test sets of the eLife and PLOS datasets. All performances are evaluated using the script provided by the organizers before the deadline for the validation sets of both datasets. The baselines are evaluated on the Codabench platform for the test set, but only the proposed model is evaluated after the challenge's deadline on the test set.

As shown in Table 2, the proposed model achieves better performance for relevance metrics with the best scores of ROUGE-1 (0.4681), ROUGE-2 (0.131), ROUGE-L (0.4475), and BERTScore (0.8404). It also attains the best scores for CLI (10.4805) and LENS (63.4962) metrics on the validation set of eLife. Meanwhile, the validation set of PLOS shows significantly improved performance for the readability metrics FKGL (14.1426), CLI (15.0911), and LENS (52.9523) compared to the listed baselines. Additionally, the BERTScore relevance metric performed well, scoring 0.858.

As observed from Table 3, the proposed model demonstrates superior performance in relevance metrics, achieving the best scores for ROUGE-1 (0.4635), ROUGE-2 (0.1228), ROUGE-L (0.4428), and BERTScore (0.8411). It also achieves top scores for CLI (10.9776) and LENS (65.7387) metrics on the eLife test set. In the PLOS test set, the model significantly improves readability metrics, with FKGL (13.8401), CLI (15.1084), and LENS (52.9811) scores surpassing those of the baselines. Additionally, the relevance metrics for the PLOS test set show notable improvements, with ROUGE-1 (0.4396), ROUGE-L (0.3988), and a strong BERTScore of 0.8578.

5 Conclusion and Future Scope

A two-stage fine-tuning framework combining general-purpose BART and PEGASUS LLMs with the biomedical-specific BioBART LLM showcased satisfactory performance for generating lay summaries of biomedical scientific articles. In the first phase, BART and PEGSUS were fine-tuned on the training data of eLife and PLOS datasets, while in the second phase, BioBART was fine-tuned on the merged and deduplicated lay summary generated in the first phase. All hyperparameters were set using the validation set. This framework achieved promising results for relevance and readability metrics but at the cost of marginally lower performance for factuality metrics. In the future, different combinations of domain-specific LLMs can be employed, along with language simplification techniques, to generate high-quality lay summaries for non-experts, optimizing scores for relevance, readability, and factuality. metrics. In the future, different combinations of domain-specific LLMs can be employed, along with language simplification techniques, to generate high-quality lay summaries for non-experts, optimizing scores for relevance, readability, and factuality.

6 Limitations

Adapting LLMs to new domains requires substantial fine-tuning, which may not always transfer knowledge effectively across the target domain. In the proposed model, fine-tuning of LLMs in two consecutive stages results in high computational cost and time. A significant problem is the length limitation associated with these LLMs. Although important information often resides at the beginning of scientific articles, the need to restrict input

Model	**eLife**									
	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE	FKGL	DCRS	CLI	LENS	AlignScore	SummaC
T5	0.386	0.0964	0.3742	0.8211	9.232	**6.5566**	11.7579	24.1684	0.3513	**0.7091**
BART	0.4442	0.1086	0.4145	0.8357	13.5909	10.169	13.5704	43.9951	**0.7752**	0.7033
PEGASUS	0.4068	0.1006	0.3919	0.8291	**9.5149**	6.5603	10.8472	42.6606	0.717	0.6209
BIOBART	0.4325	0.109	0.3669	0.8277	19.1071	9.7971	13.1439	27.72	0.5935	0.5149
Ours	**0.4681**	**0.131**	**0.4475**	**0.8404**	10.2465	7.8174	**10.4805**	**63.4962**	0.6264	0.5263

Model	**PLOS**									
	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE	FKGL	DCRS	CLI	LENS	AlignScore	SummaC
T5	0.3402	0.0973	0.3113	0.8304	15.469	**9.2281**	16.4606	35.0077	0.8612	0.7057
BART	0.4483	0.1456	0.4053	0.8565	14.3844	11.8589	15.7053	49.3006	0.8766	0.8281
PEGASUS	**0.455**	**0.1561**	**0.4123**	0.8579	14.6704	11.5939	16.2672	49.6905	0.8055	**0.8736**
BIOBART	0.4323	0.1479	0.3893	0.85	14.2208	12.07	15.9739	51.6555	**0.8991**	0.8396
Ours	0.4525	0.1458	0.4109	**0.858**	**14.1426**	11.4252	**15.0911**	**52.9523**	0.801	0.7152

Table 2: Results on the validation sets of eLife and PLOS datasets.

Model	**eLife**									
	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE	FKGL	DCRS	CLI	LENS	AlignScore	SummaC
T5	0.2614	0.0453	0.2404	0.8015	16.0014	7.6351	16.4143	26.2303	**0.9157**	0.691
BART	0.4479	0.1054	0.4169	0.8385	13.9387	10.3756	14.3257	49.2375	0.8131	**0.7296**
PEGASUS	0.3987	0.096	0.383	0.8295	**9.8937**	**6.5202**	11.1935	44.0906	0.7398	0.6401
BIOBART	0.4343	0.1043	0.3589	0.8308	20.131	9.9165	13.7818	30.8928	0.638	0.5341
Ours	**0.4635**	**0.1228**	**0.4428**	**0.8411**	10.3915	7.7965	**10.9776**	**65.7387**	0.6409	0.5443

Model	**PLOS**									
	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE	FKGL	DCRS	CLI	LENS	AlignScore	SummaC
T5	0.3316	0.0995	0.3034	0.8324	14.8563	**9.1379**	16.4704	34.5145	0.8647	0.6981
BART	0.4317	0.1376	0.391	0.8556	14.4732	11.8719	15.9178	49.9099	0.8876	0.8423
PEGASUS	0.4363	**0.1439**	0.3932	0.851	14.8366	11.6496	16.473	13.8482	0.8116	**0.8695**
BIOBART	0.4248	0.142	0.3839	0.8508	14.2641	12.0685	16.338	52.5272	**0.9066**	0.8366
Ours	**0.4396**	0.1405	**0.3988**	**0.8578**	**13.8401**	11.3222	**15.1084**	**52.9811**	0.8183	0.7273

Table 3: Results on the test sets of eLife and PLOS datasets.

length and truncate the rest can lead to a potential loss of information, ultimately affecting the quality of the generated summary.

References

Irfan Al-Hussaini, Austin Wu, and Cassie Mitchell. 2023. Pathology dynamics at biolaysumm: the trade-off between readability, relevance, and factuality in lay summarization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 592–601.

Jennifer Bishop, Qianqian Xie, and Sophia Ananiadou. 2022. Gencomparesum: a hybrid unsupervised summarization method using salience. In *Proceedings of the 21st workshop on biomedical language processing*, pages 220–240.

Xiaoyan Cai, Sen Liu, Libin Yang, Yan Lu, Jintao Zhao, Dinggang Shen, and Tianming Liu. 2022. Covidsum: A linguistically enriched scibert-based summarization model for covid-19 scientific papers. *Journal of Biomedical Informatics*, 127:103999.

Chao-Yi Chen, Jen-Hao Yang, and Lung-Hao Lee. 2023. Ncuee-nlp at biolaysumm task 2: Readability-controlled summarization of biomedical articles using the primera models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 586–591.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 468–477.

Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024. Overview of the biolaysumm 2024 shared task on lay summarization of biomedical research articles. In *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Bangkok, Thailand. Association for Computational Linguistics*.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages

10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Akanksha Karotia and Seba Susan. 2022. Pre-training meets clustering: A hybrid extractive multi-document summarization model. In *International Conference on Hybrid Intelligent Systems*, pages 532–542. Springer.

Akanksha Karotia and Seba Susan. 2023. Covsumm: an unsupervised transformer-cum-graph-based hybrid document summarization model for cord-19. *The Journal of Supercomputing*, 79(14):16328–16350.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Quancheng Liu, Xiheng Ren, and VG Vinod Vydiswaran. 2023. Lhs712ee at biolaysumm 2023: Using bart and led to summarize biomedical research articles. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 620–624.

Phuc Phan, Tri Tran, and Hai-Long Trieu. 2023. Vbd-nlp at biolaysumm task 1: Explicit and implicit key information selection for lay summarization on biomedical long documents. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 574–578.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Venkat praneeth Reddy, Pinnapu Reddy Harshavardhan Reddy, Karanam Sai Sumedh, and Raksha Sharma. 2023. Iitr at biolaysumm task 1: lay summarization of biomedical articles using transformers. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 625–628.

Mong Yuan Sim, Xiang Dai, Maciej Rybinski, and Sarvnaz Karimi. 2023. Csiro data61 team at biolaysumm task 1: Lay summarisation of biomedical research articles using generative models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 629–635.

Oisín Turbitt, Robert Bevan, and Mouhamad Aboshokor. 2023. Mdc at biolaysumm task 1: Evaluating gpt models for biomedical lay summarization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 611–619.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

eLife: The eLife dataset contains research papers associated with lay summaries written by domain experts. The train, validation, and test sets consist of 4346, 241, and 142 articles, respectively. The average word count for lay summaries across all data splits ranges between 382-400, while for articles, it ranges from 8900-10201. Similarly, the average sentence count for lay summaries and articles ranges between 18-19 and 382-583, respectively. The minimum and maximum word counts for the train, validation, and test sets for lay summaries are (177, 686), (234, 672), and (244, 642), respectively. For articles, the minimum and maximum word counts for the train, validation, and test sets are (324, 28696), (3408, 23048), and (2492, 16880), respectively.

PLOS: This includes research papers and their corresponding lay summaries from domain experts. The dataset is divided into training, validation, and test sets with 24773, 1376, and 142 articles, respectively. Lay summaries have an average word count between 180 and 195 across all splits, whereas the articles range from 6742 to 6754 words. The average sentence length for lay summaries is 8; for articles, it is between 298 - 311 sentences. For lay summaries, the minimum and maximum word counts are 4 and 511 for the training set, 55 and 384 for the validation set, and 16 and 293 for the test set. Articles have word count ranges of 748 to 26643 for the training set, 751 to 20423 for the validation set, and 1587 to 18477 for the test set.

Data	Elife			PLOS		
	Train	Validation	Test	Train	Validation	Test
Number of samples	4346	241	142	24773	1376	142
Avg. word count (LS)	382	390	400	195	195	180
Avg. sentence count (LS)	18	18	19	8	8	8
Max. word count (LS)	686	672	642	511	384	293
Min. word count (LS)	177	234	244	4	55	16
Avg. word count (A)	10200	10021	8909	6754	6742	6939
Avg. sentence count (A)	382	583	445	299	298	311
Max. word count (A)	28696	23048	16880	26643	20423	18477
Min. word count (A)	324	3408	2492	748	751	1587

Table 1: Detailed statistics and analysis of eLife and PLOS datasets, where LS stands for Lay Summary and A stands for Article.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE	FKGL	DCRS	CLI	LENS	AlignScore	SummaC
T5	0.3631	0.0969	0.3428	0.8258	12.3505	7.8924	14.1093	29.5881	0.6063	0.7074
BART	0.4463	0.1271	0.4099	0.8461	13.9877	11.014	14.6379	46.6479	**0.8259**	**0.7657**
PEGASUS	0.4309	0.1284	0.4021	0.8435	**12.0927**	**9.0771**	13.5572	46.1756	0.7613	0.7473
BIOBART	0.4324	0.1285	0.3781	0.8436	16.664	10.9336	14.5589	39.6878	0.7463	0.6773
Ours	**0.4603**	**0.1384**	**0.4292**	**0.8492**	12.1946	9.6213	**12.7858**	**58.2243**	0.7137	0.6208

Table 2: Average scores achieved for eLife and PLOS on the validation set.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE	FKGL	DCRS	CLI	LENS	AlignScore	SummaC
T5	0.2965	0.0724	0.2719	0.817	15.4289	8.3865	16.4424	30.3724	**0.8902**	0.6946
BART	0.4398	0.1215	0.404	0.8471	14.206	11.1238	15.1218	49.5737	0.8504	**0.786**
PEGASUS	0.4175	0.12	0.3881	0.8403	12.3652	**9.0849**	13.8333	28.9694	0.7757	0.7548
BIOBART	0.4296	0.1232	0.3714	0.8458	17.1976	10.9925	15.0599	41.71	0.7723	0.6854
Ours	**0.4516**	**0.1317**	**0.4208**	**0.8495**	**12.1158**	9.5594	**13.043**	**59.3599**	0.7296	0.6358

Table 3: Average scores achieved for eLife and PLOS on the test set.