AUTH at BioLaySumm 2024: Bringing Scientific Content to Kids

Loukritia Stefanou and **Tatiana Passali** and **Grigorios Tsoumakas**
School of Informatics, Aristotle University of Thessaloniki
{loukritia,scpassali,greg}@csd.auth.gr

Abstract

The BioLaySumm 2024 shared task at the ACL 2024 BioNLP workshop aims to transform biomedical research articles into lay summaries suitable for a broad audience, including children. We utilize the BioBART model, designed for the biomedical sector, to convert complex scientific data into clear, concise summaries. Our dataset, which includes a range of scientific abstracts, enables us to address the diverse information needs of our audience. This focus ensures that our summaries are accessible to both general and younger lay audience. Additionally, we employ specialized tokens and augmentation techniques to optimize the model's performance. Our methodology proved effective, earning us the 7th rank on the final leaderboard out of 57 participants.

1 Introduction

Lay summarization (i.e. summarization for non-expert audiences) helps make scientific literature understandable to non-experts. It simplifies complex technical information into clear, easy-to-understand language, promoting public understanding of research findings. The significance of lay summarization, which bridges the gap between scientific insights and public knowledge, has been increasingly recognized (Chandrasekaran et al., 2020; Goldsack et al., 2023).

To address the challenge of dense technical language in biomedical research papers, the BioLaySumm 2024 shared task (Goldsack et al., 2024) at the ACL 2024 BioNLP workshop focuses on turning biomedical research into lay summaries. These summaries need to be accurate and understandable to a broad audience, since they serve an important role in informing the public about scientific developments and avoiding the spread of misinformation. The shared task is based on data from two sources of biomedical articles: eLife and PLOS (Goldsack et al., 2022).

Our approach to this shared task is based on the BioBART-v2 model(Yuan et al., 2022), which has been demonstrated to be highly effective in summarizing biomedical content. On top of it, we employ a controllable generation technique using special tokens, in order to exploit in a single model the data from both eLife and PLOS and at the same time during inference align the produced summaries with the unique characteristics of the corresponding source, such as length, readability and level of abstraction.

In addition, towards improving the simplicity of the produced summaries, we employed augmentation to extend the eLife and PLOS data. We identified the most complex lay summaries in these datasets, and paired their source abstracts with summaries produced by GPT-4 (OpenAI et al.). To make it produce simple lay summaries, we used in-context learning providing to it examples of scientific articles targeted at children from the Science Journal for Kids.

2 The SJK Dataset

Science Journal for Kids (SJK) is a non-profit organization based in Texas that is dedicated to presenting scientific research in a manner that is accessible and appealing to children. They achieve this goal by digitally publishing on their web site[1], adaptations of scientific papers that are made to be kid-friendly. The adaptation process undertaken by SJK involves using common vocabulary and relatable examples, and then further validating and refining the adapted content for educational use. This process ensures that the content is not only accessible but also retains the educational value of the original scientific research.

The kid-friendly articles in the SJK web site are available in PDF format. To assemble the SJK dataset, we collected the PDFs of the articles and

[1] https://sciencejournalforkids.org/

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 793–803
August 16, 2024. ©2024 Association for Computational Linguistics

extracted their content. From this content we kept the abstracts and the links at the references section, pointing to the original scientific papers. We extracted DOI numbers from these links and attempted to retrieve the abstracts via the Semantic Scholar API. However, since the API often returned empty abstracts, we resorted also to an extensive scraping process on specific pages with a particular format to get the abstracts. This was not always feasible due to restrictions on scraping from certain sites, necessitating manual addition of links and abstracts to the dataset, highlighting the challenging nature of the data collection process.

We crawled the SJK web site on January 22, 2024. We initially collected all 306 articles from the SJK web site for potential future work, ensuring we had a comprehensive dataset to expand or refine as needed. From these, we eventually selected 285 articles based on their formatting suitability. Older versions had a completely different format that the scraping process couldn't recognize because it was based on the newer versions. Formatting issues included instances where the text from the abstract was cut off when scraped due to the two-column format or where scraping couldn't find the reference. We manually conducted additional checks to append missing text and locate references in these cases. Additionally, we prioritized articles that included references. Besides ensuring the credibility of the content, this allowed us to pair the kid-friendly articles with the corresponding scientific articles that inspired them.

Our final dataset[2] comprises 300 pairs, each consisting of an abstract from a scientific paper and its corresponding abstract from the children's article. We focused on abstracts because they provided comprehensive information suitable for our lay summarization task. For each article, we sourced the corresponding abstract from the first reference cited in the children's articles published by SJK, and in 25 cases, also the second reference, which are the original academic papers that the SJK articles are based on.

The articles were intentionally curated to encompass a wide array of subjects, specifically chosen to attract the scientific curiosity of young learners across disciplines such as biology, chemistry, and more. Table 1 illustrates the diversity of topics covered by both all SJK articles and our final dataset.

Table 1: Number of articles in the SJK web site and in our collection per category. Note that some articles belong to multiple categories.

Category	Ours	SJK
Biodiversity-And-Conservation	83	85
Health-And-Medicine	77	81
Biology	63	70
Energy-And-Climate	57	57
Social-Science	51	57
Water-Resources	48	48
Pollution	30	30
Food-And-Agriculture	25	26
Technology	20	23
Paleoscience	16	18
Chemistry	13	13
Physical Science	2	18

3 Our Approach

3.1 Model

Our approach employs the BioBART-v2 model. BioBART-v2 introduces significant improvements in its training methodology to advance its capabilities in the biomedical field. Unlike its precursor, which utilized a general-domain vocabulary, BioBART-v2 incorporates a specialized cross-domain vocabulary, substantially enlarging its lexicon to 85,401 tokens. This expansion is derived from Domain-Adaptive Pre-Training (DAPT) (Gururangan et al., 2020) on the PubMed abstracts corpus, resulting in a rich dataset that provides a more targeted pretraining foundation.

The construction of this vocabulary was achieved by merging the original BART's general-domain vocabulary with newly generated biomedical tokens, specifically designed from the PubMed corpus. This process yielded 60,000 additional tokens that, when combined with the existing vocabulary, boosted the model's capabilities for biomedical literature.

BioBART-v2, with its 400 million parameters, balances model complexity and computational feasibility, making it suitable for both research and practical applications. Fine-tuning it is straightforward due to its architecture, allowing targeted training on biomedical tasks with minimal computational resources. This adaptability makes it ideal for various applications in the biomedical domain, from information extraction to summarization.

[2] The dataset will be publicly available after notification of acceptance.

3.2 Data

We fine-tuned BioBART-v2 on the union of the two biomedical datasets offered by BioLaySumm 2024, i.e. PLOS and eLife. Preliminary experiments using a different model for each dataset led to inferior results. We used only the abstracts of the academic articles as sources. When properly written, the abstract of an article serves as a concise summary of the whole article, containing all the aspects needed for *translating* it into lay language. In addition, these abstracts align well with the content that was used for the pre-training of BioBART. Details about each of the two datasets follow.

The PLOS dataset comprises 26,291 articles from five peer-reviewed journals of the Public Library of Science (PLOS) publisher, covering diverse fields such as Biology, Computational Biology, Genetics, Pathogens, and Neglected Tropical Diseases. The lay summaries in this dataset are written by the authors of the articles themselves. These summaries typically range from 150 to 200 words in length. The dataset is divided into 24,773 training, 1,376 validation, and 142 testing articles.

The eLife dataset, contains 4,729 articles from the eLife biomedical journal, covering a wide array of topics in life sciences and medicine. In contrast to PLOS, eLife features lay summaries produced collaboratively by expert editors and the original authors. This collaboration resulted in summaries that are longer, more abstractive, and generally more readable. The dataset is divided into 4,346 documents for training, 241 for validation, and 142 for testing.

3.3 Data Augmentation

To improve the readability of the produced summaries, we extended the provided eLife and PLOS datasets by using GPT-4 to rewrite lay summaries of high complexity. To identify such summaries, we used three readability metrics: Flesch-Kincaid Grade Level (FKGL) (Flesch, 1948), Dale-Chall Readability Score (DCRS) (Chall and Dale, 1995), and Coleman-Liau Index (CLI) (Coleman and Liau, 1975). These metrics offer quantitative assessments of text complexity and measure the accessibility of the content across various age groups. Table 2 provides an interpretation of the FKGL metrics, illustrating how different score ranges correspond to reading and school levels. The DCRS and CLI scores similarly provide insights into the readability and complexity of the text. This approach is in alignment with the evaluation criteria of the BioLaySumm 2024 shared task.

Flesch-Kincaid Score	**Reading Level**
0 - 3	Kindergarten
3 - 6	Elementary
6 - 9	Middle School
9 - 12	High School
12 - 15	College
15 - 18	Post-grad

Table 2: Flesch-Kincaid Grade Level (FKGL) Metrics Interpretation

We specifically targeted the top 200 summaries from each of the eLife and PLOS datasets based on their highest FKGL scores, with the aim of simplifying them to reach the level of middle school students. In the eLife dataset, these summaries had average scores of FKGL 10.74, DCRS 12.39, and CLI 8.91, which correspond to high school and college reading levels. The PLOS dataset exhibited even higher complexity, with average scores of FKGL 14.73, DCRS 15.75, and CLI 10.86, aligning with college and post-graduate reading levels. These summaries, characterized by complex sentence structures and a high density of abstract ideas, were selected for augmentation to enhance their readability and accessibility for a middle school audience.

The DCRS and CLI metrics further support the interpretation of text complexity. DCRS scores above 10 indicate a higher level of text difficulty, often requiring college-level comprehension. Similarly, CLI scores, which reflect the number of characters per word and words per sentence, indicate higher complexity with scores above 8. The high DCRS and CLI scores of the selected summaries ensured that we focused on content that was particularly challenging, necessitating simplification for better accessibility.

To refine the summaries for children, we utilized the GPT-4 model via the OpenAI API, employing in-context learning via few-shot prompts to guide our augmentation pipeline. In particular, two randomly selected kids-friendly abstracts from the SJK dataset were used as examples during the augmentation process. These examples acted as guidelines, ensuring that the adapted summaries met the desired standards of simplicity. Additionally, the prompt asked to simplify the language and make it more accessible. An example of the prompt used for this purpose is illustrated below:

> *"You're explaining scientific concepts to a kid who's curious to learn. Keep all the important facts, but use easier words that are easier for kids to understand. Here are two examples of how to do it:"*
>
> *1. [A random kid-friendly abstract from the SJK dataset],*
> *2. [Another random kid-friendly abstract from the SJK dataset]*

Tables 3 and 4 present the mean scores of the original and augmented summaries. These scores demonstrate significant improvements in the readability of the augmented versions of the lay summaries.

Table 3: Mean readability scores for General (Targeted 200) and Kids summaries in the eLife dataset.

Category	FKGL	DCRS	CLI
Original	10.74	12.39	8.91
Augmented	7.90	8.99	7.33

Table 4: Mean readability scores for General (Targeted 200) and Kids summaries in the PLOS dataset.

Summary Type	FKGL	DCRS	CLI
Original	14.73	15.75	10.86
Augmented	8.57	9.07	7.50

3.4 Controllable Generation

Our methodology employs special tokens in the source abstracts to achieve two distinct controllable generation goals: i) adapt the produced summary towards the specific style of either of the two datasets, ii) guide the summary generation towards increased readability.

For the first goal, we use special tokens `<elife>` and `<plos>` to differentiate between the two datasets, as from the analysis in Sections 3.2 we know that expert-written eLife summaries are longer and more readable. For the second goal, we use special tokens `<general_lay_summary>` and `<kids_lay_summary>` to differentiate abstracts that are paired with original lay summaries from abstracts that are paired with augmented lay summaries adapted for children.

During training, we prepend each PLOS abstract with the `<plos>` tag and each eLife abstract with the `<elife>` tag. In addition, we prepend the augmented abstracts with the `<kids_lay_summary>` tag, while the rest of the abstracts are prepended with the `<general_lay_summary>` tag.

During inference, we again prepend each PLOS and eLife abstract with the corresponding `<plos>` and `<elife>` tags, while we experiment with including one or both of the `<general_lay_summary>` and `<kids_lay_summary>` tags to control the readability of the produced lay summary. Our final submission included both tags, as this led to the best results in the validation sets.

4 Results and Discussion

This section presents and discuss the results on the validation datasets provided by eLife and PLOS.

4.1 Experimental Setup

The fine-tuning process of BioBART-v2 was conducted using the Amazon Web Services (AWS) cloud platform. We utilized AWS S3 for storing model steps and output data. The fine-tuning tasks were executed on Amazon SageMaker, using a `p3.2xlarge` instance equipped with NVIDIA Tesla V100 GPU. More details on the experimental setup can be found in Appendix A.1.

We evaluated all the models using a combination of metrics to assess the *relevance*, *readability*, and *factuality* of the generated summaries, based on the BioLaySumm 2024 shared task. The relevance of the summaries was measured by metrics including ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L) (Lin, 2004), and BERTScore (Zhang* et al., 2019) to assess how well the content matched the original articles. Readability was evaluated through metrics such as the Flesch-Kincaid Grade Level, Dale-Chall Readability Score, Coleman-Liau Index, and LENS. Factuality was verified using AlignScore (Align S.) (Zha et al., 2023) and SummaC (Laban et al., 2021) to check the accuracy of the information presented in the summaries.

Our experimental results include the following variants:

- **Baseline:** This refers to the model's performance when trained using only the original scientific content of the eLife and PLOS datasets, without any additional data or special tokens.

Table 5: Experimental results on PLOS and eLife datasets.

Step	Approach	R-1	R-2	R-L	BertScore	FKGL	DCRS	CLI	LENS	Align S.	SummaC
					PLOS						
300	Baseline	0.488	0.171	0.449	0.850	14.149	11.142	14.370	73.706	0.778	0.636
300	Sp. Token	**0.494**	**0.173**	**0.454**	**0.865**	14.430	11.321	14.552	**74.978**	**0.790**	**0.647**
300	Sp. Token + Aug.	0.490	0.167	0.451	0.864	**13.839**	**10.914**	**13.336**	72.242	0.789	0.651
					eLife						
400	Baseline	0.479	0.133	0.453	0.838	10.979	8.813	11.541	72.445	0.622	0.539
400	Sp. Token	0.488	0.135	0.458	0.852	11.152	8.991	11.745	73.182	0.634	0.547
400	Sp. Token + Aug.	**0.491**	**0.135**	**0.462**	**0.851**	**10.636**	**8.750**	**11.284**	**73.707**	**0.640**	**0.548**
					Combined						
300+400	Sp. Token + Aug.	0.491	0.151	0.457	0.857	12.237	9.832	12.310	72.974	0.714	0.599

- **Sp. Token:** Represents the performance of the model when it has been added to with special tokens. This configuration does not include any augmented data.

- **Sp. Token + Augmented:** This configuration includes the use of special tokens, as mentioned above, along with the data augmentation strategy.

4.2 Results

This subsection highlights the summaries produced by our models at their best-performing steps during the competition. These results demonstrate the effectiveness of our specialized configurations, including the use of special tokens and augmented data, aimed at improving both the accessibility and accuracy of the summaries.

We detail the performance metrics for the PLOS and eLife, illustrating significant improvements in readability as a result of our modeling efforts, as shown in Table 5. Two different checkpoints were selected for the final summaries of the eLife and PLOS to optimize generation in line with the unique characteristics and challenges of each dataset. The chosen checkpoints reflect points where the model achieved an optimal balance between relevance, readability, and factual accuracy specific to each dataset. A more detailed analysis regarding each of the relevance, readability and factuality metrics along with detailed plots illustrating the different training steps can be found inA.2.

The use of special tokens consistently improved relevance scores across both datasets, indicating their effectiveness in helping the model understand the context and semantics better. Without special tokens, the model's relevance scores were notably lower, showing that it struggled to capture the essential details of the scientific content. This pattern was observed in both the eLife and PLOS datasets, highlighting the critical role of special tokens in enhancing the model's performance.

5 Conclusion

Our approach to the BioLaySumm 2024 shared task showcases BioBART's ability to simplify complex biomedical research articles into accessible lay summaries. By fine-tuning BioBART with specialized tokens and data augmentation techniques, we generated readable summaries for specific audiences, including younger readers.

A key aspect of our methodology was the use of specialized tokens to precisely control the characteristics of each dataset and audience. Additionally, we enriched our dataset with kid-friendly content from the Science Journal for Kids, enabling us to produce summaries that effectively bridge the gap between scientific complexity and public understanding. Our experimental results highlight the effectiveness of our approach, especially in improving the readability and relevance of the summaries.

While our methodology significantly improved readability and relevance, maintaining factual accuracy remains a challenge. Ensuring the factuality of lay summaries is especially critical in the biomedical field, where accuracy is important.

Our model achieved an 7th place out of 55 participants, demonstrating its validity in managing diverse and complex summarization tasks. This achievement shows the potential of our techniques in making scientific knowledge more accessible to the general public and children.

6 Limitations

In this work, we employed the BioBART model with specialized tokens and data augmentation techniques to generate lay summaries of biomedical research articles. While our approach improved the readability and relevance of the summaries, we did not explicitly analyze the factual accuracy of the generated summaries, which remains a critical issue in the biomedical domain. The introduction of augmented data, while beneficial for readability, sometimes compromised content relevance and factual accuracy. To improve the quality of our training examples, future research could integrate factuality metrics to evaluate the accuracy of generated summaries and use post-editing techniques or human review to remove inaccurate content.

Acknowledgements

AWS resources were provided by the National Infrastructures for Research and Technology GRNET and funded by the EU Recovery and Resiliency Facility.

References

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books. Google-Books-ID: 2nbuAAAAMAAJ.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview and insights from the shared tasks at scholarly document processing 2020: CL-SciSumm, LaySumm and LongSumm. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 214–224, Online. Association for Computational Linguistics.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. 60(2):283–284. Place: US Publisher: American Psychological Association.

R. Flesch. 1948. A new readability yardstick. 32(3):221–233.

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.

Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024. Overview of the biolaysumm 2024 shared task on the 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.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Preprint*, arxiv:2111.09525 [cs].

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 technical report. *Preprint*, arxiv:2303.08774 [cs].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. *Preprint*, arxiv:2305.16739 [cs].

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT.

A Appendix

A.1 Experimental Setup

Here, we present additional details regarding the experimental setup.

A.1.1 Distribution of text lengths

Firstly, as part of our configuration, we determined that the maximum input length would be set at 400 words based on the distribution of text lengths across our datasets, as shown in Table 6. This table provides the 95th percentile of text lengths and the percentage of texts that are 400 words or fewer, demonstrating that the chosen maximum input length effectively covers the majority of the data.

Table 6: Distribution of text lengths in the validation set.

Dataset	95th Perc. Length (words)	% $\leq$ 400 words
eLife Abstracts	186.09	100.0
PLOS Abstracts	368.00	97.02

A.1.2 Training Configuration

We fine-tuned and configured parameters using the Hugging Face Transformers library (Wolf et al., 2020) to ensure maximum efficiency. After a limited preliminary exploration of hyperparameter values on the validation sets of eLife and PLOS, we established the most effective settings. We set the learning rate at 1×10^{-5} to balance the speed and stability of the learning process.

We chose a batch size of 4 for both training and evaluation to optimize GPU memory usage. The model underwent training over 15 epochs, with evaluations and model savings every 50 steps to consistently monitor and evaluate progress.

A key component was the use of gradient accumulation, where we applied 64 steps. This method effectively increases the batch size to 256 (4 times 64), allowing us to handle larger batches and stabilize the training dynamics without requiring additional memory.

Thus, the number of data samples processed at each checkpoint can be determined by the following formula:

$$\begin{aligned}\text{Train Batch Size} \times \text{Gradient Acc. Steps} \times \text{Save Steps}\\ = 4 \times 64 \times 50\\ = 12,800\end{aligned}$$

A.2 Detailed Analysis and Training Plots

Here, we provide a more detailed analysis regarding the effectiveness of each approach across different steps in terms of relevance, readability, and factuality. For the sake of presentation clarity, we selected three indicative training checkpoints for detailed examination, which summarize the whole training process. We used different numbers of steps for the eLife and PLOS datasets to better present the key outcomes for each dataset.

A.2.1 Relevance

The relevance of the generated summaries is measured using ROUGE scores. As shown in Tables 7 and 8, the relevance for the eLife dataset significantly improved with training, reflecting the increasing ROUGE scores. This improvement suggests that the eLife dataset, which includes longer, and more readable lay summaries written by expert editors, provides new and varied content that the model effectively learns from during training.

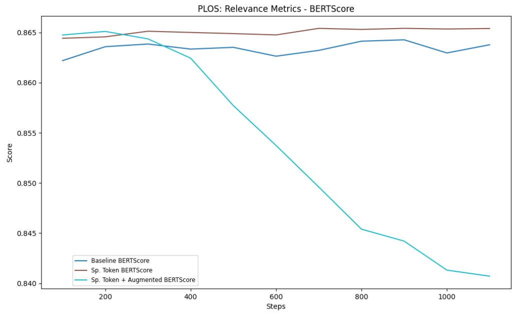

Figure 1: BERTScore relevance metric for PLOS articles.

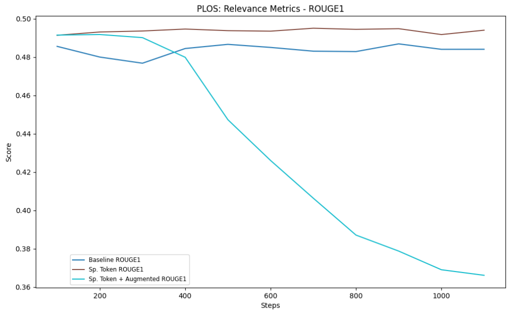

Figure 2: ROUGE-1 relevance metric for PLOS articles.

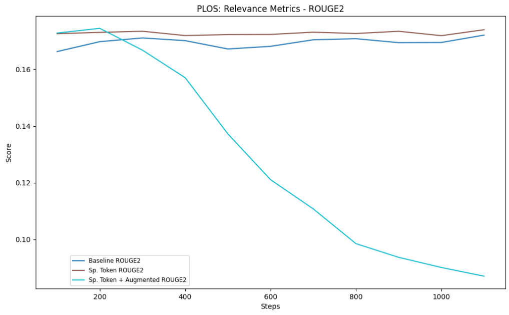

Figure 3: ROUGE-2 relevance metric for PLOS articles.

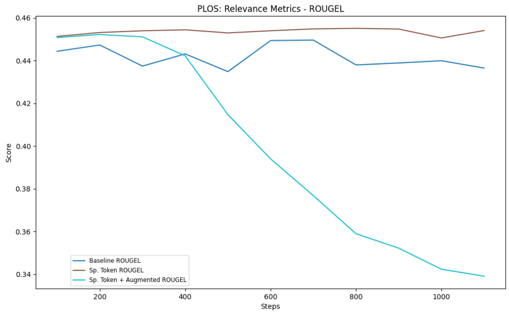

Figure 4: ROUGE-L relevance metric for PLOS articles.

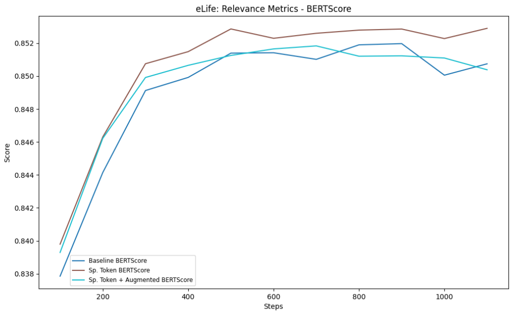

Figure 5: BERTScore relevance metric for eLife articles.

Table 7: Metrics for the eLife dataset at selected steps

Step	Approach	R-1	R-2	R-L	BERT	FKGL	DCRS	CLI	LENS	Align	SummaC
100	Baseline	0.376	0.090	0.338	0.838	13.29	10.44	14.39	56.70	0.760	0.625
100	Sp. Token	0.380	0.093	0.351	0.840	13.58	10.65	14.82	58.66	0.761	0.646
100	Sp. Token + Aug.	0.373	0.088	0.345	0.839	13.52	10.66	14.91	59.08	0.766	0.649
500	Baseline	0.488	0.134	0.451	0.850	10.63	8.75	11.42	71.53	0.641	0.543
500	Sp. Token	0.493	0.138	0.463	0.853	10.96	8.95	11.55	73.96	0.642	0.556
500	Sp. Token + Aug.	0.494	0.136	0.465	0.851	10.15	8.56	10.69	76.15	0.608	0.540
900	Baseline	0.493	0.137	0.453	0.851	10.55	8.72	10.89	72.58	0.619	0.519
900	Sp. Token	0.501	0.141	0.471	0.853	10.75	8.82	11.30	74.91	0.620	0.527
900	Sp. Token + Aug.	0.497	0.138	0.468	0.851	9.64	8.24	10.02	78.22	0.583	0.546

Table 8: Metrics for the PLOS dataset at selected steps

Step	Approach	R-1	R-2	R-L	BERT	FKGL	DCRS	CLI	LENS	Align	SummaC
100	Baseline	0.476	0.167	0.437	0.864	14.03	10.96	14.17	72.12	0.782	0.632
100	Sp. Token	0.491	0.173	0.451	0.864	14.51	11.32	14.41	74.89	0.784	0.643
100	Sp. Token + Aug.	0.491	0.173	0.451	0.865	14.50	11.29	14.32	74.95	0.783	0.641
300	Baseline	0.480	0.169	0.447	0.863	14.27	10.99	14.28	73.48	0.788	0.632
300	Sp. Token	0.494	0.173	0.454	0.865	14.43	11.32	14.55	74.98	0.790	0.647
300	Sp. Token + Aug.	0.490	0.167	0.451	0.864	13.84	10.91	13.34	72.24	0.789	0.651
600	Baseline	0.479	0.170	0.438	0.863	14.21	10.94	14.26	72.78	0.794	0.637
600	Sp. Token	0.493	0.172	0.454	0.865	14.61	11.37	14.73	75.10	0.796	0.654
600	Sp. Token + Aug.	0.426	0.121	0.394	0.854	11.36	9.16	10.44	58.34	0.791	0.657

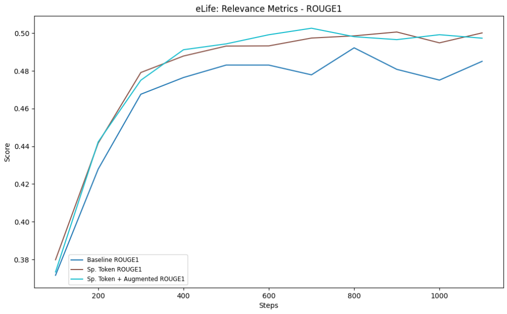

Figure 6: ROUGE-1 relevance metric for eLife articles.

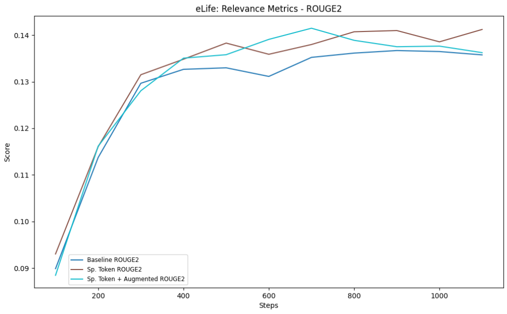

Figure 7: ROUGE-2 relevance metric for eLife articles.

For the PLOS dataset, however, the relevance did not show significant improvement with training, suggesting that the model might have already been exposed to similar data during its initial training on the PubMed archive, where PLOS articles are included (for example here: PubMed archive).

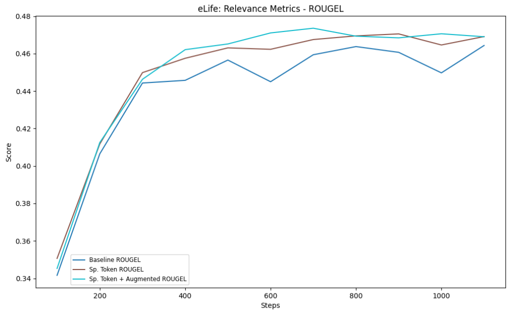

Figure 8: ROUGE-L relevance metric for eLife articles.

Additionally, the introduction of augmented data led to a decline in relevance at later steps, suggesting that the diversity brought by augmentation may complicate content relevance when the model has already encountered similar datasets.

A.2.2 Readability

Readability generally improved across successive training steps, as indicated by the FKGL, CLI, DCRS, and LENS scores in Tables 7 and 8. For the eLife dataset, the use of special tokens, along with training on new, unseen data, helped reduce complexity, making the summaries easier to understand. This consistent improvement is likely due to the nature of eLife's longer, more detailed, and editor-written summaries. Special tokens, and also augmented data, further aided this process by helping the model capture and organize the relevant contextual information more effectively.

For the PLOS dataset, while training did not significantly affect relevance or factuality, it did improve readability. This indicates that even if the model had seen similar data before, the fine-tuning process still contributed to producing more readable summaries. Augmented data helped improve readability scores, simplifying the text.

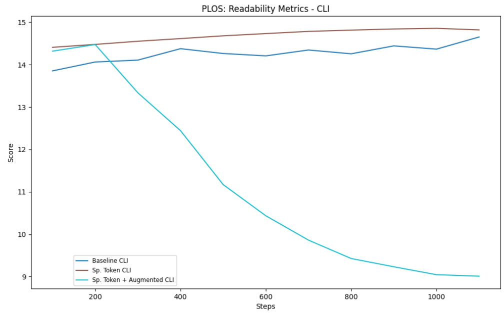

Figure 9: Coleman-Liau Index readability metric for PLOS articles.

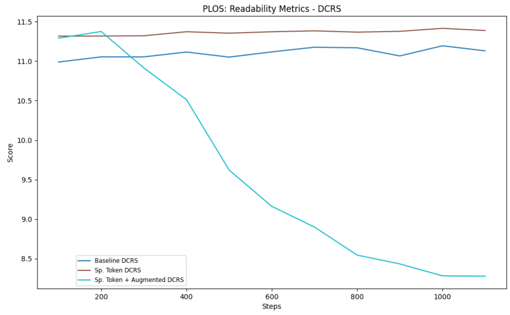

Figure 10: DCRS readability metric for PLOS articles.

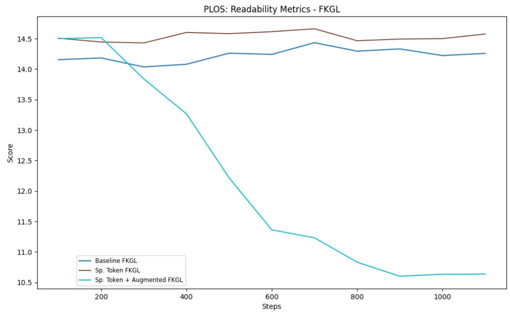

Figure 11: Flesch-Kincaid Grade Level (FKGL) readability metric for PLOS articles.

A.2.3 Factuality Metrics

Factuality metrics reveal a complex pattern of performance. For the eLife dataset, while factuality

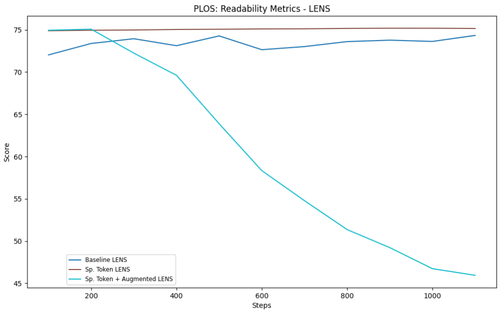

Figure 12: LENS readability metric for PLOS articles.

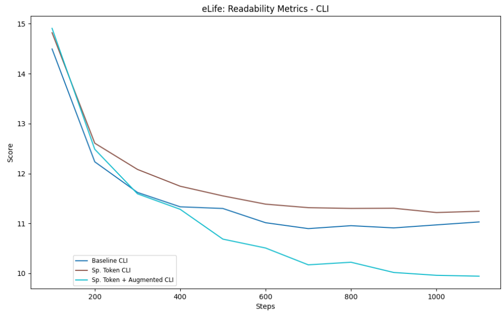

Figure 13: Coleman-Liau Index readability metric for eLife articles.

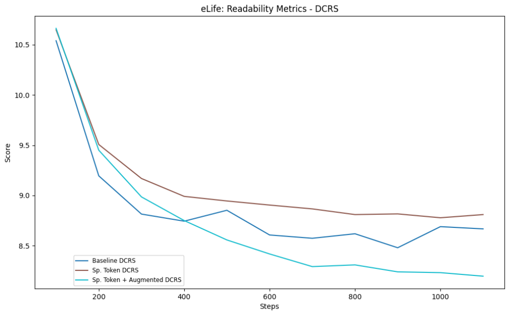

Figure 14: DCRS readability metric for eLife articles.

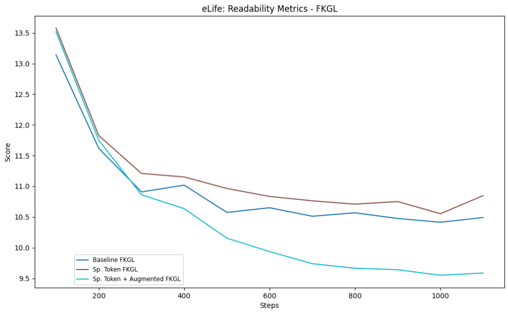

Figure 15: Flesch-Kincaid Grade Level (FKGL) readability metric for eLife articles.

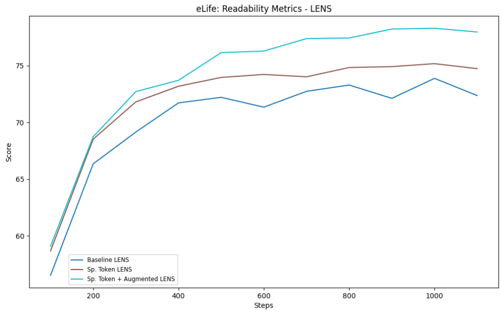

Figure 16: LENS readability metric for eLife articles.

scores showed some improvement with training, the introduction of augmented data sometimes led to a decline in factuality, especially in later steps. This suggests challenges in maintaining accuracy when introducing more diverse training data, particularly for a dataset that is initially more abstractive.

For the PLOS dataset, factuality scores did not consistently improve with training and decreased in later steps, particularly with the introduction of augmented data. This suggests that adding more diverse data did not help maintain factual accuracy and may have introduced complexity.

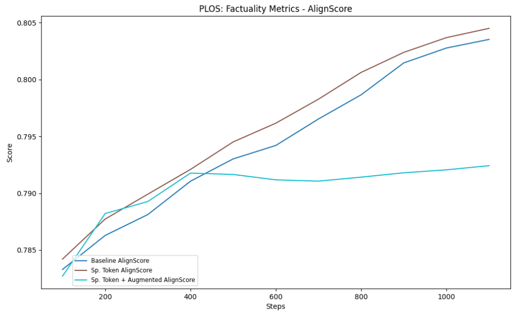

Figure 17: Alignment Score factuality metric for PLOS articles.

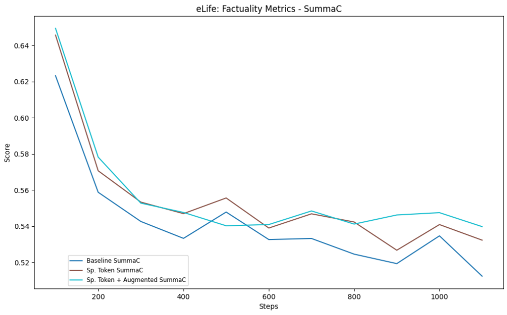

Figure 18: SummaC factuality metric for eLife articles.