MetninOzU at BioLaySumm2025: Text Summarization with Reverse Data Augmentation and Injecting Salient Sentences

Egecan Çelik Evgin[1], İlknur Karadeniz[1,2], Olcay Taner Yıldız[1,2]
[1]Department of Artificial Intelligence and Data Engineering, Özyeğin University, Türkiye
[2]Department of Computer Science, Özyeğin University, Türkiye
egecan.evgin@ozu.edu.tr,
{ilknur.karadeniz, olcay.yildiz}@ozyegin.edu.tr

Abstract

In this paper, we present our approach to the BioLaySumm 2025 Shared Task on lay summarization of biomedical research articles, which was conducted as part of the BioNLP Workshop 2025. This marks the third edition of the BioLaySumm Shared Task (Goldsack et al., 2023, 2024; Xiao et al., 2025). The aim of the task is to create lay summaries from scientific articles to improve accessibility for a non-expert audience. To this end, we applied preprocessing techniques to clean and standardize the input texts, and fine-tuned Qwen2.5 (Team, 2024; Team) and Qwen3-based language models (Yang et al., 2025; Team, 2025) for the summarization task. For abstract-based fine-tuning, we investigated whether we can insert salient sentences from the main article into the summary to enrich the input. We also curated a dataset of child-friendly articles with corresponding gold-standard summaries and used large language models to rewrite them into more complex scientific variants to augment our training data with more examples.

1 Introduction

Interdisciplinary collaboration is a major challenge, especially in the biomedical field, where the number of scientific publications is increasing rapidly and the language used is often highly technical. This complexity poses significant obstacles not only for researchers from other disciplines, but also for the general public, making it difficult to access and understand new scientific findings. One promising solution to this problem is the inclusion of lay summaries in biomedical research articles. These summaries serve as a bridge between specialized content and a broader audience, allowing students, interdisciplinary researchers, and laypeople to better understand and engage with biomedical advances. The BioLaySumm 2025 Shared Task aims to improve automated systems for generating summaries of biomedical research articles. The focus is on producing summaries that are factually accurate, accessible to non-specialists and faithful to the original scientific content, thus supporting the wider dissemination and understanding of biomedical knowledge.

Previously, Bao et al. (2024) investigated simple preprocessing techniques such as hard truncation and text fragmentation and showed that large language models can produce effective lay summaries of biomedical texts even without complex pipelines. Stefanou et al. (2024) developed a child-friendly summarization method by fine-tuning biomedical models to simplified summaries. They used specialized tokens and data augmentation to improve accessibility for younger readers, using training data from the Science Journal for Kids (Science Journal for Kids, 2024). Modi and Karthikeyan (2024) showed that minimal preprocessing of summaries such as removing parenthetical content can significantly improve LLM performance in lay biomedical summarization. You et al. (2024) applied an extract-then-summarize strategy and tuned GPT-3.5 (OpenAI, 2023) on salient sentences to achieve strong relevance and overall performance. These studies show how different approaches, from basic cleanup to structured extraction, aim to make biomedical lay summaries clearer and more accessible.

2 Datasets

The task included two datasets, PLOS and eLife (Goldsack et al. (2024) (Goldsack et al., 2022) (Luo et al., 2022)). *PLOS* is the largest dataset derived from the Public Library of Science, comprising 24,773 training instances and 1,376 for validation, while the *eLife* dataset was derived from the peer-reviewed *eLife* journal and contains 4,346 instances for training and 241 for validation. The test data used for evaluation consisted of examples from both sources and was kept hidden by the organizers.

BioNLP 2025 Shared Tasks, pages 179–184

3 Methodology

We investigated the pre-processing of full texts, the use of summaries and full articles for lay summarization, the generation of synthetic data from child-friendly texts with LLMs, and the extraction of key phrases by clustering.

3.1 Preprocessing

Before fine-tuning, we evaluated the performance of zero-shot and few-shot models using raw text input. The models tested include DeepSeek-Qwen (Lyu et al., 2025) and Qwen2.5 (Yang et al., 2024) with either full articles or abstracts provided as input. Building on the principles of PoA (Preprocessing over Abstract) from Modi and Karthikeyan (2024), we introduce a preprocessing step PoWA (Preprocessing over Whole Article) that improves the performance in both zero-shot and few-shot scenarios. PoWA involves removing all content enclosed in square, round or curly brackets from the input text including those in the training and test sets.

As with many systems submitted in previous years, our initial strategy focused on using only the abstract as input text for the summary. However, due to the varying lengths of the abstracts, we adopted a consistent approach by selecting the first 10 sentences from each abstract. The sentence boundaries were determined by splitting on periods, and applied uniformly to both the training and validation sentences. Unless otherwise specified (e.g. the condition "Full test" in Table 1), we only used the first 10 sentences of each test instance during the tests. This ensured comparability between different models and configurations.

3.2 Reverse Data Augmentation

Following the approach of Stefanou et al. (2024), we adopted a fine-tuning enhancement strategy that incorporates external data. Specifically, we used Frontiers for Young Minds(Frontiers for Young Minds, 2024), a child-friendly branch of the Frontiers journal series(Frontiers, 2024), which features simplified scientific articles written for young audiences. We collected 373 articles from the 'Human Health' section using a web scraping script built with the Selenium library (Selenium Project, 2025). Each article includes an abstract and spans approximately 500–1000 words. Designed for readers aged 8 to 12, these texts employ low FKGL (Flesch-Kincaid Grade Level) language (Flesch, 1975), with accompanying abstracts that provide even more simplified summaries. Each abstract was treated as a golden summary, resulting in a data set with two columns: Article and Summary. However, since both the article texts and their summaries were already simplified, the resulting pairs did not reflect the input-output complexity of the task. To address this gap, we used the DeepSeek-R1-Distill-Qwen-32B model (DeepSeek-AI, 2024) to rewrite the simplified articles in a more scientific tone, following the method described by DeepSeek-AI (2024). We used the following prompt: "Rewrite the given text so that it is more scientific and suitable for publication." The generation was limited to 1024 tokens with a temperature of 0.01 and a repetition penalty of 1.2. As mentioned in DeepSeek-AI (2024), DeepSeek models often produce internal thoughts before generating the final output. To address this, we extract the content following the </think> tag, along with minimal pre- and post-processing to format the results.

The gold summaries from Frontiers for Young Minds typically had FKGL scores between 8 and 10 (Flesch, 1975), and were notably shorter than the summaries found in the eLife and PLOS training sets (Task, 2025a,b). To address this length and complexity mismatch, we incorporated a curriculum learning strategy (Bengio et al., 2009), which is discussed further in Section 4.4 on model fine-tuning.

3.3 Injecting Salient Sentences

Using only the abstract to summarize an entire article was found to be insufficient. To improve this and build on strategies observed in our earlier literature review, we appended key sentences from the full text to the end of each abstract. To process sentences beyond the initial 10 in each article, we developed a function that encodes these sentences using the all-MiniLM-L6-v2 model (Wang et al., 2020), which is accessible via the Hugging Face repository (Reimers and Gurevych, 2021).

We trained a K-Means clustering model with $k = 3$ on encoded sentence representations to identify the salient content (Lloyd, 1982). A sentence closest to each centroid was selected, resulting in three sentences in total, which were then appended to the end of the article's abstract. Transformers and Scikit-learn libraries were used for this phase (Wolf et al., 2020; Pedregosa et al., 2011).

3.4 Model Fine-tuning

First, fine-tuning was performed only on the abstract and lay summary pairs using the Qwen2.5: 1.5B and Qwen2.5: 3B3B models (Team, 2024; Team), prompted with a very short instruction: *"Summarize the following:"* The Qwen2.5 models were fine-tuned using low-rank adaptation (LoRA) (Hu et al., 2021).

For Qwen3 models (Yang et al., 2025; Team, 2025), we applied LoRA for parameter-efficient fine-tuning, using a rank of 8, a scaling factor of 16, and a dropout rate of 0.05. Adaptation was limited to the `q_proj` and `v_proj` attention layers, without any bias terms, under a causal language modeling setup (Hu et al., 2021; Dettmers et al., 2023).

After preprocessing steps such as trimming, salient sentence injection, curriculum learning, and adding Frontiers for Young Minds articles, the data was converted into ChatML format (OpenAI, 2023) and used for fine-tuning.

Training hyperparameters were slightly adjusted based on the dataset. For eLife, we fine-tuned the model for 3 epochs with a learning rate of 1×10^{-4} and 6 gradient accumulation steps. For PLOS, we used 2 epochs, a higher learning rate of 1.5×10^{-4}, and 8 accumulation steps. For other datasets, we set the learning rate to 1.25×10^{-4}, trained for 2 epochs, and used 7 accumulation steps. These values were chosen after a few initial trials to balance training time and performance. All models were trained with a per-device batch size of 2 and FP16 precision using Hugging Face Transformers and PEFT libraries (Wolf et al., 2020; Dettmers et al., 2023).

We applied curriculum learning (Bengio et al., 2009), which is presented in Table 1 with "Aug" label, in which 373 articles from *Frontiers for Young Minds* were placed at the beginning of the training dataset (Frontiers for Young Minds, 2024),as explained in Section 3.2. The remaining articles were then sorted by word count in ascending order, resulting in a training sequence that gradually progressed from simpler to more complex texts. In the Salient Sentence Injection strategy (see Section 3.3), the three most important sentences following the abstract were added to it, and fine-tuning was done on this updated version of the dataset. The part marked as *Full Text* in Table 1 refers to the evaluation of the two 142-entry test sets *without any trimming*, prior to fine-tuning. The experiment labeled as "Post Processing" in the same table refers to the action taken after fine-tuning, as described in Section 3.5.

3.5 Post-processing for Readability

To slightly reduce the FKGL (Flesch, 1975) score of the summaries generated by the fine-tuned LLMs, a post-processing step was applied. Using the DeepSeek-R1-Distill-Qwen-32B model (DeepSeek-AI, 2024) in a zero-shot setting, we prompted it with: "Reduce the FKGL score of the text. Simplify while preserving the scientific content" DeepSeek-AI (2024). As in Section 3.2, post-processing was also applied to the outputs of the DeepSeek model (DeepSeek-AI, 2024). In most experiments, additional steps and alternative prompts were needed due to the model frequently disrupting the structure of the article.

4 Experimental Setup

The training was performed on an NVIDIA A100 GPU (Corporation, 2020) provided by Google Colaboratory (Bisong, 2019). Several automatic metrics to measure relevance were used for evaluation, with a focus on comparing system output with human-written references. ROUGE (Lin, 2004) evaluates recall by measuring the overlap of n-grams between the generated text and the reference text. BLEU (Papineni et al., 2002) focuses on the precision of the n-grams and applies a penalty for brevity to prevent overly short outputs. METEOR (Banerjee and Lavie, 2005) considers synonym matching, stemming and word order, balances precision and recall, and penalizes disjointed output. BERTScore (Zhang et al., 2020) captures semantic similarity by calculating cosine similarity between contextualized token embeddings from models such as BERT (Devlin et al., 2019), enabling a deeper evaluation of meaning beyond surface-level overlaps.

The Flesch-Kincaid Grade Level (FKGL)(Flesch, 1975) assesses the reading difficulty of a text based on sentence length and word syllables and provides a score that corresponds to US school levels. The Coleman-Liau Index (CLI)(Coleman and Liau, 1975) provides a similar assessment of readability, but is based on the number of characters rather than the number of syllables, making it more suitable for automatic processing of digital texts. The D-Level Sentence Complexity Rating Scheme (DCRS)(Rambow

Model	ROUGE	BLEU	METEOR	BERTScore	FKGL	DCRS	CLI	LENS	AlignScore	SummaC
Qwen3:4B Trim + Aug + SSI + PostP	0.3061	5.3966	0.2555	0.8537	16.7644	11.2446	16.0117	60.0364	0.7837	0.6858
Qwen3:4B Trim + Aug + SSI + Full Test	0.2576	4.2385	0.3296	0.8493	15.0595	10.0385	15.5170	22.2383	0.9025	0.9369
Qwen3:4B Trim + Aug + SSI	0.3261	6.6388	0.2910	0.8560	16.3742	11.0955	16.9846	34.2622	0.8748	0.9195
Qwen3:4B Trim + Aug	0.3279	6.7490	0.2928	0.8560	16.3242	11.0915	16.9893	34.0978	0.8679	0.9203
Qwen3:4B Trim	0.3300	6.9466	0.2903	0.8567	16.4528	11.2157	17.0054	34.8577	0.8807	0.9203
Qwen2.5:3B Trim	0.3127	6.2905	0.3036	0.8486	14.7591	9.8484	15.4835	23.1406	0.7937	0.9172
Qwen2.5:1.5B Trim	0.3108	6.2470	0.3014	0.8484	14.8767	9.7678	15.6298	23.1043	0.8047	0.9170

Table 1: Evaluation metrics of Qwen models on various configurations. Trim: Trimming top 10 sentences of the article, Aug: Reverse data augmentation using Frontiers for Young Minds, SSI: Salient Sentence Injection, PostP: Postprocessing for lower FKGL using DeepSeek, Full Test: Full test set in the inference without trimming

et al., 2004) assesses grammatical complexity by analyzing syntactic features such as sentence structure and part-of-speech patterns. More recently, LENS(Tan et al., 2023) uses a comprehensive language model to estimate how difficult a passage is to understand, providing a neural-based alternative to traditional readability metrics.

To assess factuality, AlignScore (Jia et al., 2022) was used to determine whether the generated summary remains faithful to the content of the source. It applies a Natural Language Inference (NLI) model (Bowman et al., 2015) to assess whether each sentence in the summary is implied by the source text. Similarly, SummaC (Laban et al., 2022) checks the factual consistency between the summary and the source by applying sentence-level entailment models to ensure logical consistency.

5 Results

The Qwen2.5-1.5B and 3B (Team, 2024; Team) models were fine-tuned with LoRA, reducing the training and validation sentences to their first 10 sentences. They were then tested with zero shot on similarly trimmed test sets, and the results were surprising. After the experiments, the lowest FKGL values were observed for the two Qwen2.5 models.

The Qwen3-4B model (Yang et al., 2025; Team, 2025) was fine-tuned with LoRA, reducing the training and validation sentences to their first 10 sentences. The highest ROUGE score was observed in the scenario where only the test set was trimmed, with no data augmentation, injection of salient sentences, post-processing, or use of the full test data (labeled 'Qwen3: 4B Trim' in Table 1). With augmentation, the FKGL score decreased slightly and the METEOR score increased slightly, but ROUGE, BLEU, BERTScore and AlignScore all decreased in the Qwen3:4B Trim + Aug setting. With the addition of Salient Sentence Injection (SSI), most relevance scores decreased and AlignScore increased slightly, which is shown in Table 1 as Qwen3:4B Trim + Aug + SSI.

In the Qwen3:4B Trim + Aug + SSI + Full Test experiment, the test set without trimming was used. As a result, ROUGE and BLEU scores decreased significantly, while METEOR, AlignScore and SummaC were higher than in all other experiments. The FKGL, CLI and LENS scores also decreased, suggesting that higher factuality could be achieved in this setting.

In our comparative analysis of the different techniques, we found that data augmentation consistently improves readability, but leads to a decrease in relevance and factuality. Salient Sentence Injection led to a decrease in all three evaluation criteria. Full fine-tuning also decreased performance in relevance and readability, but scored highest in factuality. Post-processing with external LLMs performed worst overall, scoring lowest in all experiments.

6 Conclusion

In this paper, we present our participation in the BioLaySumm 2025. Our results show that the performance of the Qwen 1.5B model with low parameters was particularly promising and shows that even smaller models can be competitive if they have sufficient input data and the hyperparameters are set appropriately. With additional input data and further optimization, this model has the potential to outperform larger counterparts, especially in terms of readability. In particular, the use of untrimmed test data significantly improved factuality, on the other hand it led to a decrease in core relevance scores. This suggests that an intermediate strategy (e.g. using a higher value for the first sentences instead of first 10 sentences) might provide a better balance between factuality and relevance. Although techniques such as salient sentence injection, reverse data augmentation, and postprocessing with auxiliary LLMs did not yield the expected gains, they remain promising for future exploration.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Siyu Bao, Ruijing Zhao, Siqin Zhang, Jinghui Zhang, Weiyin Wang, and Yunian Ru. 2024. Ctyun ai at biolaysumm: Enhancing lay summaries of biomedical articles through large language models and data augmentation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 837–844, Bangkok, Thailand. Association for Computational Linguistics.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48. ACM.

Ekaba Bisong. 2019. Google colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, pages 59–64. Apress, Berkeley, CA.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642.

Meri Coleman and T L Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

NVIDIA Corporation. 2020. Nvidia a100 tensor core gpu. Accessed: 2025-05-19.

DeepSeek-AI. 2024. Deepseek-r1-distill-qwen-32b. https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B. Accessed: 2025-04-18.

DeepSeek-AI. 2024. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*. Accessed: 2025-04-18.

Tim Dettmers, Artidoro Pagnoni, Arjun Guha, and Luke Zettlemoyer. 2023. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft. Accessed: 2024-05-18.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Rudolf Flesch. 1975. Flesch–kincaid grade level. https://en.wikipedia.org/wiki/FleschâĂŞKincaid_readability_tests. Accessed: 2025-04-18.

Frontiers. 2024. Frontiers. https://www.frontiersin.org. Accessed: 2024-12-01.

Frontiers for Young Minds. 2024. Frontiers for young minds. https://kids.frontiersin.org. Accessed: 2024-12-01.

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, Toronto, Canada. Association for Computational Linguistics.

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 *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, 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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Qingxiu Jia, Qipeng Xu, Weiting Yu, Yitong Duan, Jian-Yun Nie, and Zhiyuan Liu. 2022. Alignscore: Evaluating factual consistency with contextual alignment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Philippe Laban, Florian Trummer, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for consistency evaluation. In *Transactions of the Association for Computational Linguistics*, volume 10, pages 163–177.

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

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuxuan Lyu and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Satyam Modi and T Karthikeyan. 2024. Eulerian at BioLaySumm: Preprocessing over abstract is all you need. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 826–830, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2023. Chatml prompting format. https://platform.openai.com/docs/guides/gpt/chat-completions-api. Accessed: 2025-05-19.

OpenAI. 2023. Gpt-3.5. https://platform.openai.com/docs/models/gpt-3-5. Accessed: 2024-12-01.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Owen Rambow, Lokesh Liu, Lance Johnson, Nathaniel Fillmore, and Benoit Lavoie. 2004. Summarizing multiple news articles using readability-based evaluation. In *Proceedings of the Human Language Technology Conference of the NAACL.*

Nils Reimers and Iryna Gurevych. 2021. all-minilm-l6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Accessed: 2025-05-19.

Science Journal for Kids. 2024. Science journal for kids. https://www.sciencejournalforkids.org. Accessed: 2024-12-01.

Selenium Project. 2025. Selenium webdriver. https://www.selenium.dev. Accessed: 2025-04-18.

Loukritia Stefanou, Tatiana Passali, and Grigorios Tsoumakas. 2024. AUTH at BioLaySumm 2024: Bringing scientific content to kids. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 793–803, Bangkok, Thailand. Association for Computational Linguistics.

Zhengyuan Tan, Mounica Maddela, Wei Xu, Xiaojun Wan, and Fei Wu. 2023. Lens: A learned evaluation metric for text simplification. In *Proceedings of the 2023 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).*

BioLaySumm Shared Task. 2025a. Biolaysumm2025-elife. https://huggingface.co/datasets/BioLaySumm/BioLaySumm2025-eLife. Accessed: 2025-05-19.

BioLaySumm Shared Task. 2025b. Biolaysumm2025-plos. https://huggingface.co/datasets/BioLaySumm/BioLaySumm2025-PLOS. Accessed: 2025-05-19.

Qwen Team. Qwen2.5 models on hugging face. https://huggingface.co/Qwen. Accessed: 2024-05-18.

Qwen Team. 2024. Qwen2 technical report. https://qwen.aliyun.com/. Accessed: 2024-05-18.

Qwen Team. 2025. Qwen3 models on hugging face. https://huggingface.co/Qwen. Accessed: 2025-05-19.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and Jamie Brew. 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.

Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Sophia Ananiadou, Noura Al Moubayed, Liang Zhan, William Cheung, and Chenghua Lin. 2025. Overview of the biolaysumm 2025 shared task on lay summarization of biomedical research articles and radiology reports. In *The 24nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, and 1 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388. Available at: https://arxiv.org/abs/2505.09388.

An Yang and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhiwen You, Shruthan Radhakrishna, Shufan Ming, and Halil Kilicoglu. 2024. UIUC_BioNLP at BioLaySumm: An extract-then-summarize approach augmented with Wikipedia knowledge for biomedical lay summarization. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 132–143, Bangkok, Thailand. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.