What Are You Token About? Differentiable Perturbed Top-k Token Selection for Scientific Document Summarization

Luca Ragazzi* **Paolo Italiani*** **Gianluca Moro*** **Mattia Panni**
Department of Computer Science and Engineering, University of Bologna
{l.ragazzi, paolo.italiani, gianluca.moro}@unibo.it
mattia.panni@studio.unibo.it

Abstract

Scientific document summarization aims to condense complex and long articles in both technical and plain-language terms to facilitate the accessibility and dissemination of scientific findings. Existing datasets suffer from a deficiency in source heterogeneity, as their data predominantly stem from a single common resource, hindering effective model training and generalizability. First, we introduce SciLay, a novel dataset that includes documents from multiple natural science journals with expert-authored technical and lay summaries. Second, we propose PruneePert, a new transformer-based model that incorporates a differentiable perturbed top-k encoder layer to prune irrelevant tokens in end-to-end learning. Experimental results show that our model achieves a nearly 2x speed-up compared to a state-of-the-art linear transformer, remaining comparable in effectiveness. Additional examinations underscore the importance of employing a training dataset that includes different sources to enhance the generalizability of the models. Code is available at https://github.com/disi-unibo-nlp/sci-lay.

1 Introduction

Abstractive summarization aims to meticulously condense documents by discerning and rephrasing their salient points. Although this task is relatively easy when summarizing concise texts—such as news facts (Grusky et al., 2018; Narayan et al., 2018)—synthesizing scientific articles presents formidable challenges for humans (Altmami and Menai, 2022). In fact, experts must conduct thorough reviews, encountering technical terms and formulas (Yasunaga et al., 2019), to fully grasp the foundational information contained within the text.

To mitigate the considerable time and effort required for this endeavor, scientific document summarization (SDS) emerges as an indispensable tool.

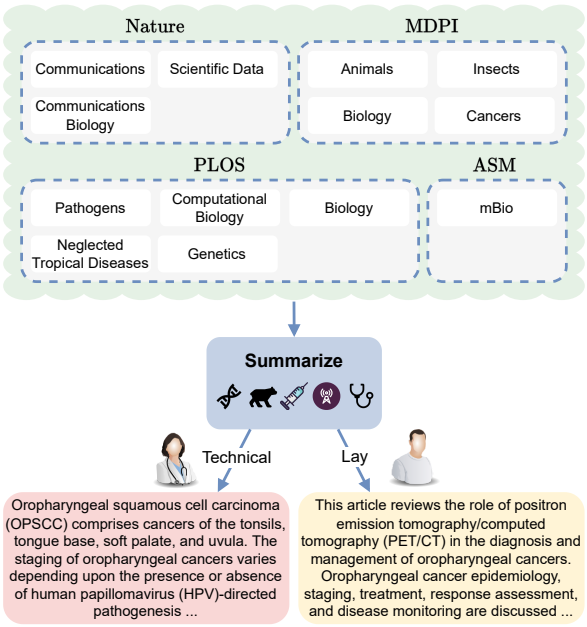

Figure 1: Overview of our SciLay benchmark.

SDS refers to the automatic production of technical and plain-language synopses from the scientific literature, including the generation of abstracts (Gharebagh et al., 2020; Frisoni et al., 2023), systematic literature reviews (Moro et al., 2022, 2023e), and journalistic reports (Dangovski et al., 2021). It plays a pivotal role in improving accessibility to the latest research findings, either by assisting experts quickly acquire the desired information or by helping the general public understand complex research topics. For example, in-domain professionals require precise technical summaries laden with specialized jargon. Conversely, non-experts typically seek more simple syntheses with layman's terminologies, complemented by contextual explanations that enhance comprehension.

To promote SDS research, several public benchmarks have been proposed over the years. However, such corpora exhibit at least one of the following limitations: (i) they are built to offer either technical (Cohan et al., 2018) or lay sum-

*These authors contribute equally to this work.

Findings of the Association for Computational Linguistics ACL 2024, pages 9427–9440
August 11-16, 2024

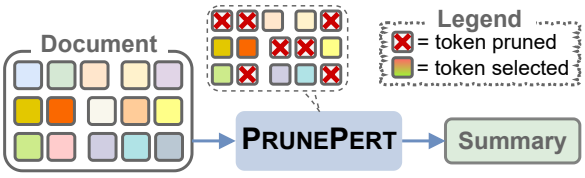

Figure 2: Overview of our PRUNEPERT model.

maries (Cardenas et al., 2023); (ii) the documents come from the same source, e.g., ELIFE (Goldsack et al., 2022), limiting model generalizability in production. In response, we introduce SCILAY (Figure 1),[1] a new open-access high-quality SDS corpus distinguished by the following characteristics: (i) it comprises author-written and expert-checked summaries of both types that carefully follow journal guidelines; (ii) it covers multiple domains from different sources (e.g., Nature Communications, PLOS Genetics, MDPI Insects, ASM mBio).

To tackle this benchmark, we explore a new model architecture to generate the summaries. Unlike previous solutions that approach this problem with conventional transformer-based pretrained language models (PLMs), we focus on addressing the following challenges. First, scientific papers exhibit an intricate and extensive structure (Kashyap et al., 2023), with summary-worthy information dispersed throughout the long input. Second, processing time and resource demand increase proportionally with the length of the input (Moro et al., 2023d), which presents a bottleneck in real-world applications for long scientific articles (Moro and Ragazzi, 2022, 2023). To this end, we propose PRUNEPERT (Figure 2), an SDS model that extends a PLM architecture with a token-pruning layer. We incorporate a differentiable perturbed top-k mechanism within the encoder stack, aiming to locate a user-defined percentage of summary-worthy input tokens in end-to-end learning. This method allows the model to process fewer input tokens, enhancing the efficiency and interpretability of PLMs.

In summary, our contributions are twofold. First, we establish a new publicly available dataset tailored for both technical and lay SDS. Second, we propose a novel model that learns to select and use only a fraction of tokens for summary generation. Through rigorous evaluations and ablation studies, we demonstrate that our proposed model can generate coherent and informative summaries

[1]The dataset is available at https://huggingface.co/datasets/disi-unibo-nlp/SciLay.

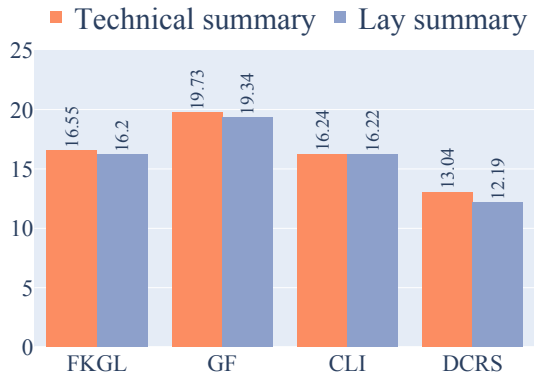

Figure 3: Average readability metrics computed on the SCILAY's test set (the lower, the more readable).

comparable to state-of-the-art baseline solutions, registering a 2x speed-up in computational time. Overall, this work contributes to the progression of SDS, offering users more efficient tools for crafting diverse forms of scientific summaries.

2 SCILAY

SCILAY is a new dataset created to summarize scientific papers for both technical and lay audiences.

Collection Unlike existing datasets in the literature (see Table 1 for a comparison), we curate a comprehensive and varied collection of journals from different publishers (see Figure 1). We scrape articles from the PubMed Central repository[2]—which archives literature from biomedical and life sciences journals—and parse the instances from XML to JSONL format. We then perform a rigorous cleaning phase, discarding samples with missing attributes and outliers based on the length of the document and summaries. Each instance of the dataset includes the full article text, lay and technical summaries, affiliated journal, keywords, Digital Object Identifier (DOI), and a unique identifier in the PubMed Central library database (PMCID).[3] We perform an 80-10-10 train/validation/test split by stratifying on the journal type, obtaining 35,026/4380/4384 instances. More details on the dataset are given in Appendix.

Readability We compare the readability of technical with their lay counterparts using the following standard metrics: Flesch-Kincaid Grade Level (FKGL, Kincaid et al., 1975), Coleman-Liau Index

[2]https://www.ncbi.nlm.nih.gov/pmc/

[3]Since not all articles on PubMed Central have lay summaries, our dataset comprises instances that originally include the source article and both a technical and lay synthesis.

Dataset	Samples	Doc	Tech Summary		Lay Summary	
		# words	# words	# sents	# words	# sents
PUBMED (Cohan et al., 2018)	133,215	2640.8	177.3	6.7	-	-
ARXIV (Cohan et al., 2018)	215,913	5282.3	237.8	8.9	-	-
LAYSUMM (Chandrasekaran et al., 2020)	572	4426.1	-	-	82.2	3.8
PLOS (Goldsack et al., 2022)	27,525	5366.7	-	-	175.6	7.8
ELIFE (Goldsack et al., 2022)	4828	7806.1	-	-	347.6	15.7
SCITECHNEWS (Cardenas et al., 2023)	2431	7570.3	-	-	216.8	7.9
SCILAY (Ours)	43,790	7530.4	239.1	8.8	145.7	5.7

Table 1: Comparison of related datasets. The number of words and sentences (sents) are averaged. For all corpora, we report values from Cardenas et al. (2023). SCILAY is the first SDS dataset with technical and lay summaries.

(CLI, Coleman and Liau, 1975), Dale-Chall Readability Score (DCRS, Dale and Chall, 1948), and Gunning Fog Index (GF, Gunning, 1952). These assessment metrics gauge the approximate number of years of education required to comprehend a given text. Lower scores signify greater readability; scores falling within the 13–16 range align with the reading proficiency expected at the college level within the US education system. Specifically, (i) FKGL assesses the total count of sentences, words, and syllables contained within the text; (ii) CLI is determined by the number of sentences, words, and characters; (iii) DCRS evaluates readability by considering the average sentence length and the presence of familiar words, using a reference table comprising the 3000 most frequently used English words; and (iv) GF calculates the average sentence length and the proportion of "hard words," defined as those containing more than two syllables. The results described in Figure 3 indicate the necessity of a college-level education to understand even the lay summary. Consistent with previous work on text simplification (Devaraj et al., 2021; Goldsack et al., 2022; Cardenas et al., 2023), the readability of both summaries is comparable; yet, we observe that lay syntheses are more readable across all metrics.

Characterization To compare the alignment of the technical and lay summaries with the input article, we calculate the extractive fragment coverage and density (Grusky et al., 2018). Technically, an extractive fragment is defined as the set of words shared between two texts. Coverage gauges the proportion of words in the summary that constitute an extractive fragment. Density employs the square of common fragment lengths, ensuring that summaries with longer common fragments receive higher values than those with more numerous but shorter common fragments. Furthermore, we compute the percentage of novel n-grams to assess the amount of information present in the summary that is not explicitly stated in the source document. Table 2 illustrates that lay summaries exhibit lower values for both density and coverage compared to their technical counterpart. In contrast, technical summaries demonstrate lower percentages of novel n-grams. These findings suggest that lay summaries manifest a reduced level of information directly replicated from the source document. In other words, lay summaries exhibit greater abstractiveness, requiring a shift in style from the source and the use of rephrasing strategies.

SCILAY	Tech Summ	Lay Summ
Coverage (↓)	0.93	0.90
Density (↓)	3.67	3.08
% novel unigrams (↑)	13.51	16.55
% novel bigrams (↑)	44.94	51.50
% novel trigrams (↑)	70.75	76.79

Table 2: Statistics of the summaries in SCILAY in terms of abstractiveness. The arrows denote the direction towards a more abstractive output.

3 PRUNEPERT

In laying the groundwork for understanding our solution (Section 3.2), we provide a conceptual preliminary with the needed foundations.

3.1 Preliminary

Given the generative nature of the tasks at hand, we employ a transformer-based encoder–decoder model (Vaswani et al., 2017), consisting of an encoder $E = \{e_1, \ldots, e_L\}$ and a decoder $D = \{d_1, \ldots, d_L\}$, each with a stack of L identical layers. $E(\cdot)$ transforms the input sequence of symbols $x = (x_1, \ldots, x_{|x|})$ into a series of continuous representations $\mathbf{h}_L = \{\mathbf{h}_{L,1}, \ldots, \mathbf{h}_{L,|x|}\}$. Then, leveraging $\mathbf{h}_L$, $D(\cdot)$ generates the output sequence $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{|y|}\}$ in an autoregressive manner.

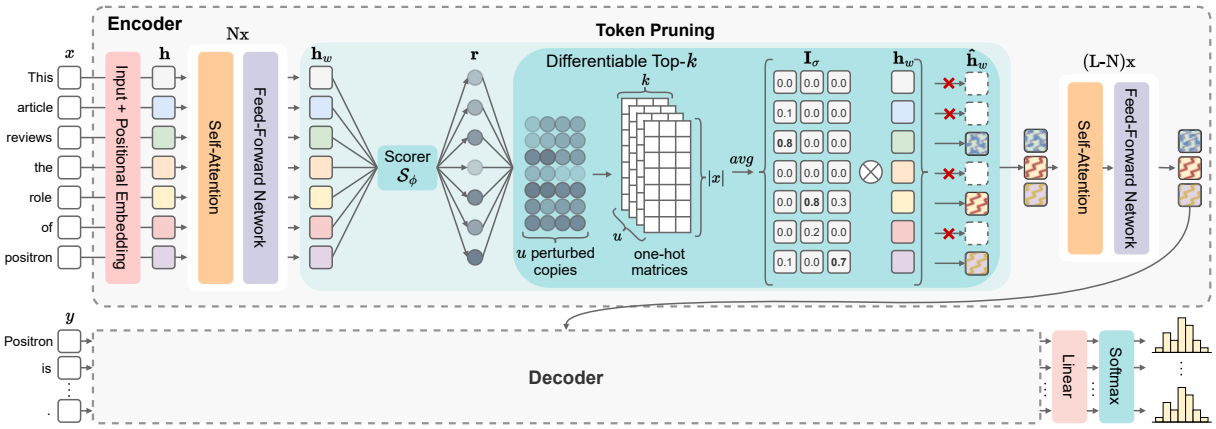

Figure 4: PRUNEPERT architecture (in this example, $k = 3$ and $u = 4$).

Specifically, every e_i layer encompasses a self-attention module (SELF-ATT) and a feedforward network module (FFN) through a residual connection and layer normalization (LN):

$$\begin{aligned} \mathbf{m}_{i-1} &= \text{LN}\left(\mathbf{h}_{i-1} + \text{SELF-ATT}\left(\mathbf{h}_{i-1}\right)\right), \\ \mathbf{h}_i &= \text{LN}\left(\mathbf{m}_{i-1} + \text{FFN}\left(\mathbf{m}_{i-1}\right)\right). \end{aligned} \tag{1}$$

In addition, $D(\cdot)$ uses a cross-attention module (CROSS-ATT) between the SELF-ATT and FFN modules, performing multi-head attention over $\mathbf{h}_L$:

$$\mathbf{c}_{i-1} = \text{LN}\left(\mathbf{m'}_{i-1} + \text{CROSS-ATT}\left(\mathbf{m'}_{i-1}, \mathbf{h}_L\right)\right), \tag{2}$$

where $\mathbf{m'}_{i-1}$ denotes the output of the SELF-ATT decoder module after being processed by LN and summed to the residual connection.

3.2 Perturbed Token Pruning

We introduce PRUNEPERT (Figure 4), a novel model that prunes the encoder at token-level granularity using a differentiable perturbed top-k selection module. To achieve our goal of end-to-end model training without the inclusion of supplementary auxiliary losses, we allow the model to pinpoint and leverage only the most significant tokens.

Model Architecture We use the transformer-based architecture to inherit the knowledge acquired by PLMs. We then integrate a scorer network $\mathcal{S}_\phi$ and a token selection module $\mathcal{T}$ into the model encoder E, where ϕ represents the trainable parameters. $\mathcal{S}_\phi$ and $\mathcal{T}$ are placed at a certain height of E, between the e_w and e_{w+1} layers, where w is a hyperparameter. Mechanically, we feed the continuous representations $\mathbf{h}_w$ and obtain relevance scores $\mathbf{r} = \mathcal{S}_\phi(\mathbf{h}_w) \in \mathbb{R}^{|x|}$ for each token in the input sequence x. Subsequently, we retain the indices of the k most salient tokens $\mathbf{i} = \mathcal{T}(\mathbf{r}) \in [1, |x|]^k$, where $\mathcal{T}$ is a discrete operator.[4] To preserve the original ordering of the hidden states, it is necessary for the $\mathbf{i}$ values to be sorted, i.e., $\mathbf{i}_1 < \mathbf{i}_z < \mathbf{i}_{z+1} < \mathbf{i}_k$. The intuition here is that if we had a large tensor of all token embeddings, we could extract the relevant ones using a single matrix multiplication. To this end, we represent each index $\mathbf{i}_z$ as the corresponding $|x|$-dimensional one-hot vector $\mathbf{I}_z$, obtaining $\mathbf{I} = \{\mathbf{I}_1, \ldots, \mathbf{I}_k\} \in \{0, 1\}^{|x| \times k}$. Thus, the pruned k hidden states are obtained as $\hat{\mathbf{h}}_w = \mathbf{I}^\top \mathbf{h}_w$.

Differentiable Top-k The token selection module $\mathcal{T}$ described above is non-differentiable. This implies that during the backpropagation of gradients through the model, no adjustments will be made to the parameters ϕ of the scorer $\mathcal{S}$. To enable differentiation through this operator, we use the perturbed maximum method (Berthet et al., 2020), whose forward pass is defined as follows:

$$\mathbf{I}_\sigma = \mathbb{E}\left[\arg\max_{\mathbf{I} \in \mathcal{C}} \langle \mathbf{I}, \mathbf{r}\mathbf{1}^\top + \sigma \mathbf{Z} \rangle\right], \tag{3}$$

where $\mathbf{r}\mathbf{1}^\top \in \mathbb{R}^{|x| \times k}$ are the scores copied k times, σ is an hyperparameter that regulates the influence of the noise, and $\mathcal{C}$ is the constraint set defined as follows:

[4] For example, $\mathbf{i}_1 = z$ implies that the first selected token is the z-th of the input x.

$$\mathcal{C} = \{\mathbf{I} \in \mathbb{R}^{|x|\times k}, \mathbf{I}_{a,b} \geq 0, \mathbf{1}^\top \mathbf{I} = \mathbf{1}, \mathbf{I}\mathbf{1} \leq \mathbf{1}, \sum_{a\in[|x|]} a\mathbf{I}_{a,k^*} < \sum_{b\in[|x|]} b\mathbf{I}_{b,k'}\ \forall k^* < k'\}. \quad (4)$$

The condition $\mathbf{1}^\top \mathbf{I} = \mathbf{1}$ ensures that each column sums up to 1, while the last constraint imposes the ordering of the $\mathbf{i}$ indices. Empirically, we sample u uniform Gaussian noises to perturb $\mathbf{r}\mathbf{1}^\top$. For each perturbed input, we run the top-k algorithm and perform the Monte-Carlo estimation of Equation 3 by averaging their results. According to Berthet et al. (2020), the Jacobian associated with the above forward pass is the following:

$$J = \mathbb{E}\left[\arg\max_{\mathbf{I}\in\mathcal{C}} \langle \mathbf{I}, \mathbf{r}\mathbf{1}^\top + \sigma\mathbf{Z}\rangle \mathbf{Z}^\mathbf{T}/\sigma\right]. \quad (5)$$

Utilizing the top-k operator on each perturbed input leads to the generation of the one-hot matrix $\mathbf{I}$. However, the average of these matrices may deviate significantly from the one-hot pattern, particularly in the initial stages of training when the scoring system is indecisive. This results in obtaining the pruned k hidden states $\hat{\mathbf{h}}_w$, forming a weighted average of the original hidden states $\mathbf{h}_w$. Accordingly, a beneficial consequence emerges wherein backpropagated gradients consider all tokens. This obviates the need to wait several iterations until the appropriate tokens are consistently sampled. At inference time, we perform hard top-k for efficiency reasons, as there is no need for u perturbed repetitions. This leads to a train–test gap; therefore, as suggested in Cordonnier et al. (2021), we linearly decrease σ to 0 at training time, so that no noise is added and the differentiable top-k operator is numerically identical to hard top-k.

4 Experiments

4.1 Experimental Setup

Hardware Environment All runs are tracked with Weights & Biases[5] and executed on a workstation with a single Nvidia GeForce RTX3090 GPU of 24 GB of dedicated memory, 64 GB of VRAM, and an Intel® Core™ i9-10900X1080 CPU @ 3.70GHz. The operating system is Ubuntu 20.04.3 LTS. To enhance consistency and portability, our development environment is built on top of a docker container with a NVIDIA image.[6]

[5] https://wandb.ai

[6] nvidia/cuda:11.3.1-devel-ubuntu20.04

Evaluation To perform a comprehensive evaluation of model performance, we address various dimensions. First, we report the F1 scores of syntactic metrics such as ROUGE-{1,2,L} (Lin, 2004), also providing $\mathcal{R}$ (Moro et al., 2023b), their variance-aware aggregated score. Second, we use the model-based metric BARTScore (Yuan et al., 2021) for semantic coverage, reporting precision, recall, and F1 values. Third, we report the average readability score across FKGL, GF, CLI, and DCRS.[7] We also conduct a thorough human evaluation. Finally, we assess the efficiency of the models by monitoring the training runtime. Additional technical details are given in Appendix.

Baselines Despite the popularity of decoder-only architectures driven by large language models (LLMs), recent findings confirm the superiority of encoder–decoder networks for text summarization (Fu et al., 2023). Therefore, we consider two widely-used encoder–decoder solutions: BART (Lewis et al., 2020), a model characterized by a denoising pretraining objective, and PEGASUS (Zhang et al., 2020), a model featuring a pretraining objective tailored to abstractive summarization. To feed sequences longer than 1024 tokens, we leverage the LSG architecture (Condevaux and Harispe, 2023) for both models, inducing $\mathcal{O}(n)$ complexity w.r.t. the input length.

4.2 Discussion

Effectiveness Table 3 presents the results on SCILAY. It is apparent that the inclusion of PRUNEPERT does not degrade the performance of PEGASUS; rather, it enhances summary quality, as evidenced by improved ROUGE and BARTScore metrics in the technical and lay summarization tasks, respectively. PRUNEPERT also contributes to the creation of more readable lay summaries, as indicated by the average readability score. Notably, solutions based on PEGASUS demonstrate superior proficiency in generating lay summaries. This phenomenon may be attributed to the increased difficulty of producing technical summaries, which demand greater length and lexical complexity.

Efficiency In addition to evaluating summarization performance, our focus lies on minimizing computational load during training. Figure 5 illustrates the substantial acceleration achieved by

[7] As typically done in the literature, we did not evaluate the readability of technical summaries since they are intended for a specialized audience, making such a study less pertinent.

Model	R-1	R-2	R-L	$\mathcal{R}$	BaS-R	BaS-P	BaS-F1	Read.
Technical Summarization								
BART	28.70	4.95	16.07	16.42	-3.363	-4.887	-4.125	-
PEGASUS	35.01	**10.06**	25.40	23.24	**-3.380**	**-2.333**	**-2.856**	-
PEGASUS + PRUNEPERT	**37.61**	9.35	**26.27**	**24.09**	-3.663	-2.403	-3.033	-
Lay Summarization								
BART	23.86	3.99	15.57	14.38	-3.299	-5.094	-4.196	16.67
PEGASUS	**40.76**	**12.69**	**28.57**	**26.98**	**-3.258**	-3.228	-3.243	16.08
PEGASUS + PRUNEPERT	38.32	10.93	27.23	25.17	-3.451	**-2.776**	**-3.113**	**15.29**

Table 3: Overall results on the SCILAY dataset. The best scores are in bold. PRUNEPERT enhances syntactic (ROUGE) and semantic (BARTScore) metrics in the technical and lay summarization tasks, respectively.

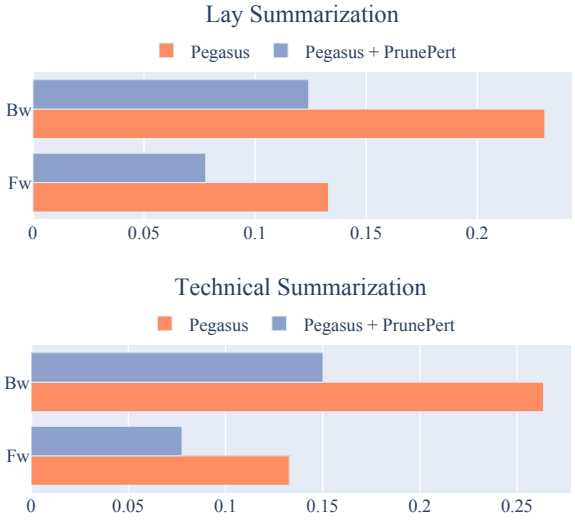

Figure 5: Average execution time (seconds) per instance throughout the forward (fw) and backward (bw) passes. PRUNEPERT achieves a speedup of up to 46%.

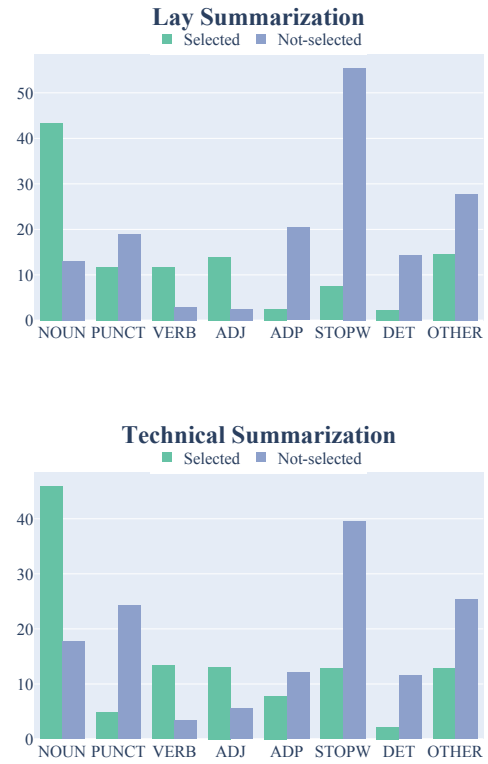

Figure 6: The average rate of POS tags (nouns, punctuation, verbs, adjectives, adpositions, stopwords, determiners, and others) processed by PRUNEPERT in both summarization tasks. We observe a tendency to preserve information stored in nouns, adjectives, and verbs.

PRUNEPERT in PEGASUS. Notably, this efficiency enhancement is evident in both the forward and backward passes of the model, resulting in a noteworthy improvement ranging from 42% to 46%. Specifically, the most significant boost in absolute terms is observed during the backward pass, which inherently demands more computational time.

Interpretability To gain a deeper understanding of how the token pruning module works, it is crucial to perform a thorough examination of both the retained and discarded tokens. Therefore, in our analysis, we specifically examine the frequency of stopwords and the following parts-of-speech (POS) tags: adjectives, nouns, verbs, adpositions, and determiners. To achieve this goal, we employ ScispaCy (Neumann et al., 2019), a library equipped with pipelines and models designed for scientific documents. Figure 6 illustrates the results of our analysis, showing the average frequency at which each POS tag and stopword appear in both selected and unselected tokens within the input article. As expected, we observe a tendency to retain elevated information content stored in nouns, adjectives, and verbs. On the contrary, other elements essential for ensuring grammatical correctness, yet offering limited informational value, are excluded. This applies both to lay and technical summarization. However,

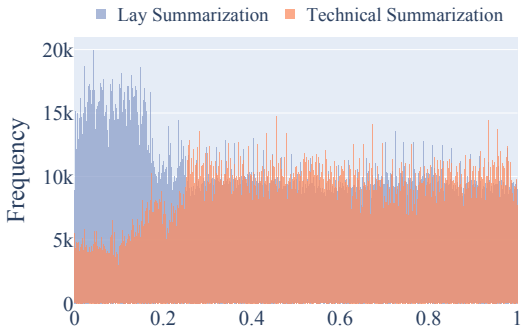

Figure 7: Normalized position distribution of tokens selected by PRUNEPERT. In lay summarization, the model tends to select a higher proportion of tokens from the first 25% of the input.

the latter is distinguished by a higher retention rate of stopwords, likely aimed at ensuring the more pronounced syntactic overlap described in Table 2.

Furthermore, in Figure 7, we examine the positional distribution of the selected tokens. Remarkably, we observe a clear differentiation between technical and lay summarization. Specifically, in lay summarization, there is a higher frequency of tokens within the first 25% of the input, whereas in technical summarization, the trend is reversed. The distribution of the last 75% tokens appears relatively uniform across both tasks, with no significant distinction. Considering the conventional role of the first section as the preamble in an article, it logically follows that, due to the need for additional contextual information in a lay summary, the selection should be particularly guided by it.

Transfer Learning To deepen our understanding of how SCILAY can contribute to SDS, we use it as a dataset for fine-tuning models, which are subsequently evaluated across benchmarks within the field. Further, our goal is to emphasize the role of source heterogeneity in enhancing the generalizability of models when applied to diverse corpora beyond their training origin. In particular, we experiment with the following two fine-tuning settings:

- **PEGASUS**$_{all}$: we train PEGASUS using 3000 instances sourced from SCILAY, ensuring that the distribution of journals within this subset is proportional to the entire dataset.
- **PEGASUS**$_{plgen}$: we train PEGASUS using 3000 instances exclusively sourced from the PLOS Genetics journal in SCILAY.

Model	R-1	R-2	R-L	BaS-F1
arXiv				
PEGASUS$_{all}$	**28.56**	**5.01**	**17.23**	**-4.437**
PEGASUS$_{plgen}$	27.42	4.53	16.33	-4.466
PubMed				
PEGASUS$_{all}$	**28.60**	**6.46**	**18.01**	**-3.934**
PEGASUS$_{plgen}$	27.76	5.82	16.96	-3.970

Table 4: Transfer learning results on external SDS datasets. The best scores are in bold. The model trained on multiple heterogeneous sources (PEGASUS$_{all}$) consistently achieves superior results.

We evaluate both configurations using 1000 instances from the test sets of the SDS datasets PUBMED and ARXIV (Cohan et al., 2018). Table 4 shows that PEGASUS$_{all}$ achieves superior performance in terms of ROUGE and BARTScore metrics in all datasets, underscoring the importance of including training instances from multiple sources.

Human Evaluation To qualitatively analyze the summaries generated by PEGASUS and PEGASUS+PRUNEPERT, we conduct a detailed human evaluation study. We randomly select 30 SCILAY's test set instances and invite three English-proficient annotators with strong NLP competencies in the biomedical domain. Upon reviewing the articles and their corresponding summaries, each evaluator assigns scores to the generated summaries using a Likert scale ranging from 1 (worst) to 5 (best), on three distinct dimensions: (i) *Recall* evaluates whether the generated summary encompasses all target contents; (ii) *Precision* examines whether the generated summary includes only the target contents without extraneous or redundant information; (iii) *Faithfulness* assesses whether the generated summary maintains factual consistency with the original text. Figure 8 presents the results of the human evaluation study, reaffirming the absence of any notable distinction among solutions.

5 Related Work

Scientific Document Summarization SDS has been a long-standing task, mainly focused on the generation of technical summaries, such as abstracts. Cohan et al. (2018) employ a hierarchical encoder to model the discourse structure and an attentive discourse-aware decoder for summary generation. An et al. (2021) integrate information from the source document and its references, using a graph-based citation model. Recently, there has

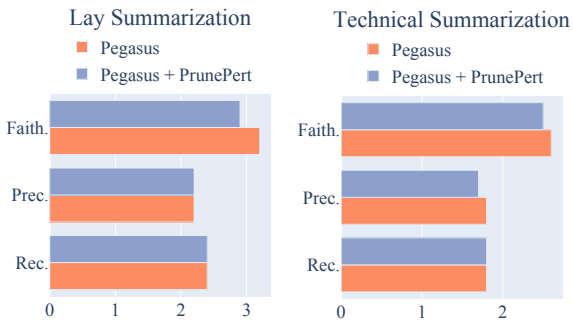

Figure 8: Human evaluation results considering faithfulness, precision, and recall, with scores ranging from 0 to 5. The outcomes among models are comparable.

been a notable shift in focus within the scientific community, expanding beyond traditional text summarization to include simplification. This shift has given rise to various datasets and methodologies to address this gap. The LaySumm task (Chandrasekaran et al., 2020) is a significant initiative that uses a corpus of 572 articles with author-generated lay summaries across various disciplines, including the Materials Science, Archaeology, Hepatology, and Artificial Intelligence journals. Zaman et al. (2020) automatically retrieve simplified and condensed versions of articles available on the Eureka Alert website, resulting in 5204 instances sourced from journals such as PLOS-ONE, Nature Communication, and Scientific Reports. Cachola et al. (2020) propose an innovative approach to extreme summarization, emphasizing core elements while omitting unnecessary methodological details. Guo et al. (2021) collected 6695 pairs of systematic reviews with their corresponding plain-language summaries from the Cochrane Database of Systematic Reviews. In alignment with our work, Goldsack et al. (2022) and Cardenas et al. (2023) provide datasets with varying readability levels. The first presents two datasets from biomedical journals (PLOS and eLife), pairing articles with manually-crafted lay summaries and abstracts. The second covers diverse domains, including Computer Science, Machine Learning, Physics, and Engineering.

Token Pruning In addition to extract-then-abstract methodologies (Moro et al., 2023c), initial attempts to mitigate the computational burden of transformers—even with linear complexity (Beltagy et al., 2020; Huang et al., 2021)—focus on the removal of non-informative tokens, mainly leveraging attention scores. PoWER-BERT (Goyal et al., 2020) uses a scoring function derived from attention scores to drop tokens based on their impact on others. Luo et al. (2022) extended this approach to Vision Transformers (ViTs), locating and dropping tokens using attention scores and fusing information from different attention heads. Yang et al. (2022) adaptively determined the quantization precision levels of the tokens (i.e., 0 bit, 4 bit, and 8 bit) based on their importance, as gauged by their attention probabilities. Despite their popularity and promising outcomes, these solutions are grounded in predefined heuristics. Hence, we advocate for an alternative research direction that involves automatically learning the token selection module. In this context, TR-BERT (Ye et al., 2021) formulates token reduction as a multi-step problem addressable with reinforcement learning, introducing an additional loss function to maximize rewards. Conversely, Cordonnier et al. (2021) opt for an end-to-end token selection approach that avoids introducing additional losses, relying on perturbation, which, while differentiable, lacks sparsity. Sander et al. (2023), using p-norm regularization, introduce the first differentiable everywhere and the sparse top-k operator. Nonetheless, it is crucial to note that these pruning-based methods are limited to encoder-only architectures, rendering them unsuitable for direct application to generative tasks.

6 Conclusion

We first introduced SciLay, a new SDS dataset designed to benchmark models in the creation of technical and lay summaries from heterogeneous sources. Second, we present PrunePert, a new PLM enriched with a token-pruning layer within the encoder stack that allows the model to select only a summary-worthy subset of input tokens for synthesis generation. Quantitative and qualitative analysis attest to the comparable performance between our method and a cutting-edge linear transformer; yet, PrunePert is notably more efficient, with an average speed-up of almost 2x. Further investigation reveals the nature of the selected tokens and the importance of having training sets covering multiple sources to enhance model generalizability.

Limitations

Our research is focused on the realm of science. Other domains like finance and law (Moro et al., 2023a) often require text simplification; yet, there is currently a dearth of publicly available technical and lay summaries for documents in these fields.

The effectiveness of our suggested PRUNEPERT model is limited by the requirements established by the specific task at hand. Engaging in the summarization of lengthy documents requires the adoption of models that feature an attention mechanism that exhibits linear scaling w.r.t. the length of the input sequence. Consequently, the performance enhancement achieved through token pruning in the forward pass is constrained compared to solutions employing models with quadratic complexity.

The restrictions of using a single 24 GB of GPU RAM and dealing with long input sequences hinder the use of LLMs, which could potentially have improved overall task performance and the effectiveness of the top-k token selection module.

Ethics Statement

While PLMs hold promise for enhancing summarization capabilities across diverse domains, it is crucial to acknowledge their limitations in ensuring the accuracy and fidelity of generated summaries. Therefore, we advocate for a cautious approach, recommending that any output produced by our proposed solution—but it applies to every generative model in the literature—undergoes manual scrutiny by domain experts before utilization for any purpose. This ethical precaution is essential to mitigate the risk of disseminating potentially erroneous or misleading information, particularly within clinical and scientific communities where accuracy and reliability are paramount.

Acknowledgements

This research is partially supported by (i) the Complementary National Plan PNC-I.1, "Research initiatives for innovative technologies and pathways in the health and welfare sector" D.D. 931 of 06/06/2022, DARE—DigitAl lifelong pRevEntion initiative, code PNC0000002, CUP B53C22006450001, (ii) the PNRR, M4C2, FAIR—Future Artificial Intelligence Research, Spoke 8 "Pervasive AI," funded by the European Commission under the NextGeneration EU program. We thank the Maggioli Group[8] for granting the Ph.D. scholarship to Luca Ragazzi and partially supporting the Ph.D. scholarship granted to Paolo Italiani.

[8] https://www.maggioli.com/who-we-are/company-profile

References

Nouf Ibrahim Altmami and Mohamed El Bachir Menai. 2022. Automatic summarization of scientific articles: A survey. *J. King Saud Univ. Comput. Inf. Sci.*, 34(4):1011–1028.

Chenxin An, Ming Zhong, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2021. Enhancing scientific papers summarization with citation graph. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12498–12506. AAAI Press.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. 2020. Learning with differentiable perturbed optimizers. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 9508–9519.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. ACL.

Ronald Cardenas, Bingsheng Yao, Dakuo Wang, and Yufang Hou. 2023. 'don't get too technical with me': A discourse structure-based framework for automatic science journalism. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1186–1202. ACL.

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. ACL.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 615–621. Association for Computational Linguistics.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Charles Condevaux and Sébastien Harispe. 2023. LSG attention: Extrapolation of pretrained transformers to long sequences. In *Advances in Knowledge Discovery and Data Mining - 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2023, Osaka, Japan, May 25-28, 2023, Proceedings, Part I*, volume 13935 of *Lecture Notes in Computer Science*, pages 443–454. Springer.

Jean-Baptiste Cordonnier, Aravindh Mahendran, Alexey Dosovitskiy, Dirk Weissenborn, Jakob Uszkoreit, and Thomas Unterthiner. 2021. Differentiable patch selection for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2351–2360. Computer Vision Foundation / IEEE.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–28.

Rumen Dangovski, Michelle Shen, Dawson Byrd, Li Jing, Desislava Tsvetkova, Preslav Nakov, and Marin Soljacic. 2021. We can explain your research in layman's terms: Towards automating science journalism at scale. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12728–12737. AAAI Press.

Ashwin Devaraj, Iain James Marshall, Byron C. Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4972–4984. ACL.

Giacomo Frisoni, Paolo Italiani, Stefano Salvatori, and Gianluca Moro. 2023. Cogito ergo summ: abstractive summarization of biomedical papers via semantic parsing graphs and consistency rewards. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12781–12789.

Zihao Fu, Wai Lam, Qian Yu, Anthony Man-Cho So, Shengding Hu, Zhiyuan Liu, and Nigel Collier. 2023. Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder. *CoRR*, abs/2304.04052.

Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross W. Filice. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1899–1905. ACL.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. *CoRR*, abs/2210.09932.

Saurabh Goyal, Anamitra Roy Choudhury, Saurabh Raje, Venkatesan Chakaravarthy, Yogish Sabharwal, and Ashish Verma. 2020. Power-bert: Accelerating bert inference via progressive word-vector elimination. In *International Conference on Machine Learning*, pages 3690–3699. PMLR.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 708–719. ACL.

R.; et al Gunning. 1952. Technique of clear writing.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 160–168. AAAI Press.

Luyang Huang, Shuyang Cao, Nikolaus Nova Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1419–1436. Association for Computational Linguistics.

Abhinav Ramesh Kashyap, Yajing Yang, and Min-Yen Kan. 2023. Scientific document processing: challenges for modern learning methods. *Int. J. Digit. Libr.*, 24(4):283–309.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

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

Kaicheng Luo, Huaxiong Li, Xianzhong Zhou, and Bing Huang. 2022. An attention-based token pruning method for vision transformers. In *Rough Sets -*

International Joint Conference, IJCRS 2022, Suzhou, China, November 11-14, 2022, Proceedings, volume 13633 of *Lecture Notes in Computer Science*, pages 274–288. Springer.

Gianluca Moro, Nicola Piscaglia, Luca Ragazzi, and Paolo Italiani. 2023a. Multi-language transfer learning for low-resource legal case summarization. *Artificial Intelligence and Law*, pages 1–29.

Gianluca Moro and Luca Ragazzi. 2022. Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11085–11093. AAAI Press.

Gianluca Moro and Luca Ragazzi. 2023. Align-then-abstract representation learning for low-resource summarization. *Neurocomputing*, 548:126356.

Gianluca Moro, Luca Ragazzi, and Lorenzo Valgimigli. 2023b. Carburacy: Summarization models tuning and comparison in eco-sustainable regimes with a novel carbon-aware accuracy. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 14417–14425. AAAI Press.

Gianluca Moro, Luca Ragazzi, and Lorenzo Valgimigli. 2023c. Graph-based abstractive summarization of extracted essential knowledge for low-resource scenarios. In *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pages 1747–1754. IOS Press.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, and Davide Freddi. 2022. Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 180–189. Association for Computational Linguistics.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, Giacomo Frisoni, Claudio Sartori, and Gustavo Marfia. 2023d. Efficient memory-enhanced transformer for long-document summarization in low-resource regimes. *Sensors*, 23(7):3542.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, and Lorenzo Molfetta. 2023e. Retrieve-and-rank end-to-end summarization of biomedical studies. In *Similarity Search and Applications - 16th International Conference, SISAP 2023, A Coruña, Spain, October 9-11, 2023, Proceedings*, volume 14289 of *Lecture Notes in Computer Science*, pages 64–78. Springer.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1797–1807. ACL.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 319–327. Association for Computational Linguistics.

Michael Eli Sander, Joan Puigcerver, Josip Djolonga, Gabriel Peyré, and Mathieu Blondel. 2023. Fast, differentiable and sparse top-k: a convex analysis perspective. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29919–29936. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tao Yang, Fei Ma, Xiaoling Li, Fangxin Liu, Yilong Zhao, Zhezhi He, and Li Jiang. 2022. Dtatrans: Leveraging dynamic token-based quantization with accuracy compensation mechanism for efficient transformer architecture. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 42(2):509–520.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R. Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7386–7393. AAAI Press.

Deming Ye, Yankai Lin, Yufei Huang, and Maosong Sun. 2021. TR-BERT: dynamic token reduction for accelerating BERT inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5798–5809. ACL.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing*

Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pages 27263–27277.

Farooq Zaman, Matthew Shardlow, Saeed-Ul Hassan, Naif Radi Aljohani, and Raheel Nawaz. 2020. HTSS: A novel hybrid text summarisation and simplification architecture. *Inf. Process. Manag.*, 57(6):102351.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Appendix

License SciLay will be released under the Creative Commons Attribution 4.0 International (CC BY) license. In fact, the documents in our dataset are open-access manuscripts sourced from various resources, each licensed under the aforementioned license, which enables free and unrestricted usage.

Dataset Our SciLay dataset has the following sources: *Nature Communications* (NC), *PLOS Genetics* (PLGEN), *PLOS Pathogens* (PLPAT), *PLOS Computational Biology* (PLCB), *PLOS Neglected Tropical Diseases* (PLNTD), *PLOS Biology* (PLB), *Biology* (B), *Communications Biology* (CB), *Scientific Data* (SD), *mBio* (MBIO), *Animals* (A), *Insects* (I), and *Cancers* (C). Table 5 shows split statistics.

In detail, NC is a multidisciplinary journal dedicated to disseminating top-tier research across a wide array of fields, e.g., biological, health, physical, chemical, and mathematical. PLGEN focuses on studies involving humans and investigations of model organisms, ranging from mice and flies to plants and bacteria. PLPAT publishes groundbreaking research that significantly improves our understanding of pathogen biology or interactions between pathogens and hosts. PLCB features works that advance our understanding of living systems on various scales through the application of computational methods, including molecules and cells, and patient populations and ecosystems. PLNTD publishes research dedicated to the pathology, epidemiology, prevention, treatment, and control of neglected tropical diseases, together with relevant contributions to public health and policy. PLB, BIO, and CB publish significant advances across biological sciences. SD shares advances from all areas of natural sciences, medicine, engineering, and social sciences. mBio reflects the vastness of the interconnected microbial world, covering symbiosis, pathogenesis, energy acquisition and conversion, climate change, geologic transformations, food and drug production, and even alterations in animal behavior. AN is exclusively dedicated to the field of animals, covering aspects of zoology and veterinary sciences. INS releases articles focusing on the biology, physiology, behavior, and management of arthropods, along with their interactions with human societies, plants, and ecosystem services. CAN covers basic, translational, and clinical studies in all types of tumors.

Training Details We list the hyperparameters used for fine-tuning and inference in Table 6.

Metrics Table 7 lists the hyperparameters of the metrics. BARTScore computes the generation probability $p(\mathbf{y}|\mathbf{x}, \theta)$ of a sequence $\mathbf{y}$ conditioned on another sequence $\mathbf{x}$, where θ are the weights of a BART model. Due to this generative approach, the evaluation dimensions vary depending on how $\mathbf{y}$ and $\mathbf{x}$ are defined. We consider the Recall, Precision and F1 settings. Technically, Recall ($\mathbf{h} \rightarrow \mathbf{r}$, $p(\mathbf{r}|\mathbf{h}, \theta)$) quantifies how easily a gold reference ($\mathbf{r}$) could be generated by the hypothesis ($\mathbf{h}$). Precision ($\mathbf{r} \rightarrow \mathbf{h}$, $p(\mathbf{h}|\mathbf{r}, \theta)$) evaluates the likelihood that the answer hypothesis could be constructed based on the gold reference. F1 ($\mathbf{h} \leftrightarrow \mathbf{r}$) is the harmonic mean of recall and precision. The aggregated score $\mathcal{R}$ is formally defined as: $\mathcal{R} = {}^{\text{avg}(r_1, r_2, r_L)}/_{1+\sigma_r^2}$, where σ_r^2 is the F1 variance. $\mathcal{R}$ penalizes model results with discrepant unigram, bigram, and longest common subsequence overlaps.

Quality Control Table 8 showcases examples of the generated technical and lay summaries, employing PEGASUS and PEGASUS+PRUNEPERT.

Split	Website	Samples	Doc	Tech Summary		Lay Summary	
			# words	# words	# sents	# words	# sents
NC	https://www.nature.com/ncomms/	6937	8906.4	168.3	6.5	47.7	2.1
PLGEN	https://journals.plos.org/plosgenetics/	3859	9554.3	257.6	9.5	195.0	7.7
PLPAT	https://journals.plos.org/plospathogens/	3650	9569.2	260.7	9.7	196.3	7.7
PLCB	https://journals.plos.org/ploscompbiol/	3237	9683.9	254.7	9.3	192.8	7.5
PLNTD	https://journals.plos.org/plosntds/	2862	6341.0	304.5	10.2	198.9	7.9
PLB	https://journals.plos.org/plosbiology/	1121	10165.1	247.3	9.1	216.7	8.0
B	https://www.mdpi.com/journal/biology	2022	5854.4	246.8	9.2	155.9	6.1
CB	https://www.nature.com/commsbio/	1084	8462.0	170.0	6.9	55.4	2.9
SD	https://www.nature.com/sdata/	907	5355.0	180.6	6.8	45.3	1.0
MBIO	https://journals.asm.org/journal/mbio	759	8403.5	246.5	9.1	145.0	5.7
A	https://www.mdpi.com/journal/animals	4887	5494.3	263.1	9.4	163.6	6.3
I	https://www.mdpi.com/journal/insects	1477	5372.2	238.0	9.1	169.7	6.9
C	https://www.mdpi.com/journal/cancers	8478	5851.8	242.4	9.3	133.4	5.3
OTHER	-	2510	7427.9	262.2	9.1	173.6	6.0

Table 5: Split statistics of the journals within SCILAY.

Hyperparameter	
Dropout rate	0.1
Learning rate	5e-5, linear scheduler
Optimizer	0.9 β_1, 0.999 β_2, 1e-2 weight decay
Batch size	1
Epochs	1
Decoding strategy	greedy search
Seed	42
k† selected tokens†	$\{0.1, \ldots, 0.5*, \ldots, 0.9\} \times \lvert x \rvert$
w encoder layer†	$\{1, 2, 3*, \ldots, 15\}$
u sampled noises†	$\{50, 100, 200*, 300, 400\}$
$\sigma^\dagger$	$\{0.05, 0.1, 0.2*, 0.3\}$

Table 6: Hyperparameters utilized for model fine-tuning and inference. † refers to values specific for PRUNEPERT. $*$ marks the final chosen value. $0.5 \times |x|$ means that we select half of the input tokens.

Metric	Bound	Hyperparameters
ROUGE	$[0, 1]$	`rouge_types=["rouge1","rouge2","rougeL"], use_aggregator=True, use_stemmer=True, metric_to_select="fmeasure"`
BARTScore	$]-\infty, 0[$	`model_checkpoint="facebook/bart-large-cnn", batch_size=4, segment_scores=False`

Table 7: Hyperparameters initialization for metrics.

Target Technical Summary	
Immunosuppressive molecules are extremely valuable prognostic biomarkers across different cancer types. However, the diversity of different immunosuppressive molecules makes it very difficult to accurately predict clinical outcomes based only on a single immunosuppressive molecule. Here, we establish a comprehensive immune scoring system (ISSGC) based on 6 immunosuppressive ligands (NECTIN2, CEACAM1, HMGB1, SIGLEC6, CD44, and CD155) using the LASSO method to improve prognostic accuracy and provide an additional selection strategy for adjuvant chemotherapy of gastric cancer (GC). The results show that ISSGC is an independent prognostic factor and a supplement of TNM stage for GC patients, and it can improve their prognosis prediction accuracy; in addition, it can distinguish GC patients with better prognosis from those with high prognostic nutritional index score; furthermore, ISSGC can also be used as a tool to select GC patients who would benefit from adjuvant chemotherapy independent of their TNM stages, MSI status and EBV status.	
PEGASUS	Tumour immunosuppression describes the suppressed host immune responses to tumour antigens, resulting in the reduction or loss of antigens on tumour cells, inhibiting the activation of immune effector cells and the decreased cell viability of cytotoxic T lymphocytes (CTLs) or natural killer cells. In recent years, an increasing number of oncologists have begun to focus their studies on anti-tumour immune responses, which might become fundamental markers in cancer immunotherapy. In recent years, despite remarkable progress in immunotherapy, there are still a considerable number of patients who cannot benefit from immunotherapy, which may be related to the immunosuppressive environment of tumours. Thus, testing the expression levels of immune checkpoints in GC patients and using valuable immune checkpoints to form a scoring system will significantly help surgeons accurately perform prognostic assessments.
PEGASUS + PRUNEPERT	The expression of immune checkpoints in gastric cancer (GC) has been shown to play a key role in the immunosuppressive environment of the tumour. However, the prognostic value of these immune checkpoints remains unclear. In this study, we evaluated the correlation between the expression of seven immune checkpoints and the prognosis of GC patients. A total of 652 patients with GC were included in the study. The patients were divided into six groups based on their GC stage, and their immune checkpoints were evaluated by immunohistochemistry (IHC). The correlation of the immune checkpoints with the prognosis was evaluated by Cox regression. The results showed that the immune checkpoint expression was significantly correlated with the GC stage (p 0.05) and with the tumour microenvironment (TME) (p > 0.05). The correlation between immune checkpoint and the TME was also significantly correlated (p = 0.05) with GC stage. In addition, the correlation of immune checkpoint with the microenvironment was significantly associated with the disease-free survival (DFS) and overall survival (OS). In conclusion, our results suggest that immune checkpoints can be used as prognostic indicators in GC.
Target Lay Summary	
Many organisms possess a true circadian clock and coordinate their activities into daily cycles. Among the simplest organisms harboring such a 24 h-clock are cyanobacteria. Interactions among three proteins, KaiA, KaiB, KaiC, and cyclic KaiC phosphorylation govern the daily rhythm from gene expression to metabolism. Thus, the control of the kaiBC gene cluster expression is important for regulating the cyanobacterial clockwork. A picture has emerged in which different KaiC phospho-states activate and inhibit kaiBC expression. However, the mechanism remains to be solved. Here, we investigated the impact of each KaiC phospho-state on kaiBC expression by introducing a model that combines the circadian transcription/translation rhythm with the KaiABC-protein oscillator. We tested 32 combinations of positive and negative transcriptional regulation. It turns out that the kaiBC expression and KaiC phosphorylation dynamics in wild type and kai mutants can only be described by one mechanism: threonine and double phosphorylated KaiC hexamers activate kaiBC expression and the unphosphorylated state suppresses it. Further, we propose that the activator-to-repressor abundance ratio very likely determines the kaiBC expression dynamics in the simulated kai mutants. Our suggested clock model can be extended by further kinetic mechanisms to gain deeper insights into the various underlying processes of circadian gene regulation.	
PEGASUS	Photoautotrophic organisms like plants and cyanobacteria are subjected to a daily light-dark rhythm and have been demonstrated to possess a self-sustained circadian clock. It consists of just three proteins KaiA, KaiB and KaiC composing a post-translational oscillator (PTO). The principal protein of the PTO is KaiC combining three intrinsic enzymatic activities, autokinase, autophosphatase and ATPase. The consensus view is that the ATPase crosstalks with the kinase/phosphatase through a structural coupling between the two rings. The three Kai proteins form stable complexes during the subjective night. KaiC forms hexamers and each KaiC monomer within the hexamer possesses two main phosphorylation sites (T432 and S431). The four forms of KaiC cycle in a stepwise fashion: unphosphorylated (U- KaiC), threonine phosphorylated (T-KaiC), both residues phosphorylated, and serine phosphorylated. In the presence of ATP, the three proteins are able to produce robust, temperature-compensated 24 h-cycles ofKaiC phosphorylation even in a test tube. In a recent study, we showed that an additional response regulator, SasA, is also suggested to drive the circadian transcriptional output without de novo expression of the kai genes. In this work, we studied whether the model can reproduce the observed dynamics of kaiBC transcription and kaiC phosphorylation.
PEGASUS + PRUNEPERT	Plants and cyanobacteria are subjected to a daily light-dark rhythm and a self-sustained circadian clock. The simplest circadian clock is a post-translational oscillator (PTO) composed of proteins KaiA, KaiB, KaiC, and the PTO KaiC hexamer. The PTO has been described as a unique circadian clock in Synechococcus elongatus. In this study, we developed a computational model of the PTO that simulates the interaction between the PTO and the kaiA and KaiC proteins. We found that the PTO regulates the expression of kaiB and kaiC, which are phosphorylated by ATPases. We also found that a combination of phosphorylated and phosphorylated -KaiCs is required for the PTO to function. This study provides a new perspective on the function of PTOs in eukaryotes.

Table 8: Qualitative examples on SCILAY.