ATLAS: Improving Lay Summarisation with Attribute-based Control

Zhihao Zhang[1], Tomas Goldsack[2], Carolina Scarton[2], Chenghua Lin[3][*]

[1]College of Economics and Management, Beijing University of Technology, China,
[2]Department of Computer Science, University of Sheffield, UK
[3]Department of Computer Science, The University of Manchester, UK

zhhzhang@bjut.edu.cn {tgoldsack1, c.scarton}@sheffield.ac.uk
chenghua.lin@manchester.ac.uk

Abstract

Automatic scientific lay summarisation aims to produce summaries of scientific articles that are comprehensible to non-expert audiences. However, previous work assumes a one-size-fits-all approach, where the content and style of the produced summary are entirely dependent on the data used to train the model. In practice, audiences with different goals and levels of expertise will have specific needs, impacting what content should appear in a lay summary and how it should be presented. Aiming to address this disparity, we propose ATLAS, a novel abstractive summarisation approach that can control various properties that contribute to the overall "layness" of the generated summary using targeted control attributes. We evaluate ATLAS on a combination of biomedical lay summarisation datasets, where it outperforms state-of-the-art baselines using both automatic and human evaluations. Additional analyses provided on the discriminatory power and emergent influence of our selected controllable attributes further attest to the effectiveness of our approach.

1 Introduction

Lay summarisation is defined as producing a summary of a scientific article that is comprehensible to non-experts (King et al., 2017). Recent work has shown that, when compared to technical abstracts, lay summaries typically are more readable (lexically and syntactically), more abstractive, and contain more background information, enabling a non-technical reader to better understand their contents (Luo et al., 2022; Cohen et al., 2021; Goldsack et al., 2023b). However, the extent to which these attributes are required within a lay summary depends largely on the specific needs of the reader. For example, a scientist from a related field will require less background information to understand an article's contents than an entirely non-technical

---
[*] Corresponding author

reader, but they might still require domain-specific jargon to be simplified or explained. Despite its obvious benefits, to our knowledge, no work has yet explored how we can enable such fine-grained control over comprehensibility-related aspects for lay summary generation.

In this paper, we propose **ATLAS** (**AT**tribute-controlled **LA**y **S**ummarization), a novel scientific summarisation approach that aims to control four attributes targeting distinct properties contributing to the overall "layness" of the generated summary, thus allowing it to cater to the specific needs of different audiences. Although recent attempts at text simplification and story generation have had success influencing the style (Martin et al., 2020; Kong et al., 2021; Sheang and Saggion, 2021) and content (Kong et al., 2021; Tang et al., 2024) of generated text using fine-grained controllable attributes, no work to our knowledge has explored this for scientific summarisation. Luo et al. (2022) recently addressed the task of readability-controlled scientific summarisation, however, this is only done at a binary level, training a model to produce either a technical or non-technical summary based on a single control token.

Our approach innovates by enabling a greater degree of controllability through the flexible handling of multiple attributes, allowing it to produce more diverse summaries and better address the specific needs of different audiences. Our results show that ATLAS outperforms state-of-the-art baselines in both automatic and human evaluations across three summary types with varying levels of technicality. Additional analyses confirm that attribute control positively influences performance, and suggest the selected control attributes are able to effectively capture the difference between technical and non-technical summaries.

2 Methodology

As discussed in §1, ATLAS aims to control four targeted attributes. We use BART-base as the base model for ATLAS as it represents the state-of-the-art benchmark in previous lay summarisation works (Guo et al., 2021; Goldsack et al., 2022).

Formally, each document $x = (x_1, x_2, ..., x_n)$ of length n, where x_i is the i-th token, is prepended with a control token sequence l such that $x = (l, \boldsymbol{x}_1, \boldsymbol{x}_2 s, ..., \boldsymbol{x}_n)$. l consists of our four selected control tokens, each of which targets distinct characteristics of the output summary that contributes to its overall comprehensibility. We describe each aspect below:

Length (L)   The length of the output summary in characters. A more lay audience may require a longer summary to aid comprehension.

Readability (R)   How easy it is to read the text. This is measured using the Flesh-Kincaid Grade Level (FKGL) metric, which estimates the reading grade level (US) required to understand the generated text based on the total number of sentences, words, and syllables present within it.

Background information (BG)   The percentage of sentences classified as containing primarily background information. Intuitively, a more lay audience will require greater levels of background information to contextualise an article.

Content word entropy (CWE)   The average entropy of content words. We hypothesise that jargon terms are likely to possess higher entropy values, thus lower average CWE is likely to be a property of more lay text. Since jargon terms are predominately nouns, we extract noun phrases as content words using *CoreNLP* library (Manning et al., 2014). We then follow Xiao et al. (2020) to calculate $I(x_i)$ entropy of a given token x_i as the negative logarithm of its generation probability $P(x_i)$, which is directly extracted from a pre-trained language model.

$$I(x_i) = -logP(x_i) \qquad (1)$$

During model training, true attribute values (as calculated on reference summaries) are used, allowing the model to learn to associate attribute values with summary properties. For all attributes, values are discretized into 10 fixed-width bins depending on their respective range in the train split (from minimum to maximum observed value), resulting in 10 unique control tokens for each attribute which are added to the vocabulary. For each attribute at test time, we use the most common bin value observed for reference summaries of the training set as attribute values.

3 Experimental Setup

Data.   We experiment on the biomedical lay summarisation datasets introduced in Goldsack et al. (2022), eLife (4.8k articles) and PLOS (27.5k articles), for which target lay summaries have been shown to contain different levels of "layness". Specifically, eLife's lay summaries have been characterized as longer, more readable, and more abstractive than those of PLOS, as well as being empirically observed to be suitable for a more lay audience. We, therefore, combine both of these datasets, allowing us to expose ATLAS to a greater variety of attribute values during training.[1] For each article in the combined dataset, we train our ATLAS to produce both the technical abstract and lay summary, using our control attributes to differentiate between them.

Evaluation.   We employ several automatic metrics to evaluate the performance of ATLAS. In line with common summarisation practice, we calculate ROUGE-1,2, and L variants (Lin, 2004) and BERTScore (Zhang et al., 2019). We also measure Dale-Chall Readability Score, a metric that estimates US grade level based on the frequency of common words.

Baselines.   To enable fair comparison, we rerun many of the baseline approaches used by Goldsack et al. (2022) (which have the abstract included in the input) on the combined datasets. Specifically, we rerun the Lead-3, Lead-K, and oracle heuristic baselines; TextRank (Mihalcea and Tarau, 2004), LexRank (Erkan and Radev, 2004), and HipoRank (Dong et al., 2021) unsupervised models; and BART and BART$_{Scaffold}$ supervised models. Here, we use the transformer-based BART base model (Lewis et al., 2020), which we fine-tune on our own datasets. BART$_{Scaffold}$ is the recreation of a model from Goldsack et al. (2022) which is trained using a binary control token (<abs> or <lay>) to produce either an abstract or lay summary for an article. This model is equivalent to that pro-

---

[1]To combine the datasets, we merge the training and validation sets. We evaluate on the test sets separately.
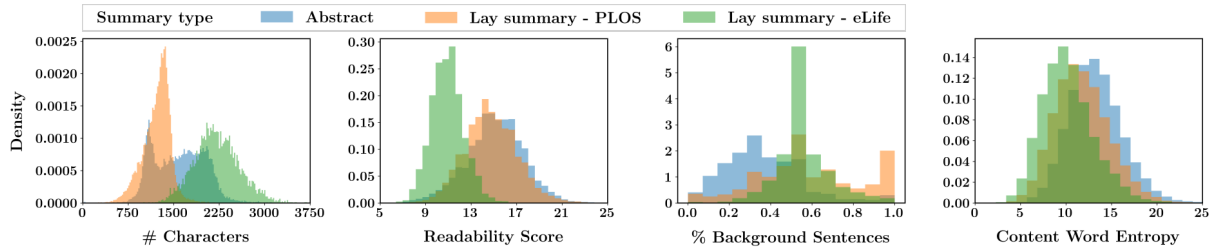
Figure 1: Visualisation of the density distributions of controllable attribute values for each summary type in the combined train split.

posed by Luo et al. (2022), the only previous work on controllable lay summarisation.[2]

Finally, we include two baselines based on Chat-GPT (3.5-turbo), so as to compare against an accessible and widely used method of controlling text generation (i.e., prompt engineering). Our first GPT baseline (GPT3.5-zs) uses the following zero-shot prompts: (i) "*Summarize the following article for an expert audience that is familiar with the technical aspects of the content*" to generate technical abstracts; (ii) "*Summarize the following article for a non-expert audience that has some familiarity with the technical aspects of the content*" to generate PLOS lay summaries, and (iii) "*Summarize the following article for a non-expert audience that has no familiarity with the technical aspects of the content*" to generate eLife lay summaries. Our second GPT baseline (GPT3.5-mdc) replicates the method of Turbitt et al. (2023), the best-performing team of the recent BioLaySumm shared task (Goldsack et al., 2023a). Based on in-context learning, this method dynamically selects the maximum number of input-output examples that fit in the context window (separated by the simple prompt "Explanation:") to generate lay summaries based on only the article abstract.

Implementation Details. As mentioned in §2, we employ BART-base as our base model. We train our ATLAS for a maximum of 5 epochs on a GeForce GTX-1080Ti GPU, retaining the checkpoint with the best average ROUGE-1/2/L score on the validation set. We set the batch size to 1 and keep the α scale factor (§2) at the default value of 0.2 from Kong et al. (2021).

For calculating control attributes, we use SciBERT (Beltagy et al., 2019) for entropy calculation, and we employ a BERT-based sequential classi-

Summary type	Precision	Recall	F1
Abstract	0.69	0.75	0.72
eLife-Lay	0.71	0.71	0.71
PLOS-Lay	0.73	0.66	0.71

Table 1: Classifier performance for 3-way classification between summary types on the combined test set.

fier (Cohan et al., 2019) trained on the PubMed-RTC dataset (Dernoncourt and Lee, 2017) for background sentence classification (as described in Goldsack et al. (2022)). We compute the FKGL readability score using the `textstat` package.

4 Experimental Results

Discriminatory ability of control attributes. To validate the ability of our controllable attributes to distinguish between different summary types, we plot the distribution of attribute values for each type in Figure 1. The figure suggests that, in combination, the attributes are able to capture characteristic differences between summary types, as instances in which two summary types share a similar distribution for one attribute can typically be separated by other attributes.[3]

To further evidence this, we use the training set to train a simple logistic regression classifier, using only the attribute values of the reference summaries as features, to discriminate between reference summary types. The test set results in Table 1 show that all summary types are classified with an F1-score above 0.7, attesting to the discriminatory power of our control attributes.

Summarisation performance. Table 2 presents the performance of ATLAS and baseline models using automatic metrics on the test sets of PLOS

---

[2]The original code for Luo et al. (2022) is not yet available at the time of writing and their results are reported on a different dataset and thus are not comparable.

[3]E.g., PLOS lay summaries and abstracts have similar readability distributions but differ in their comprehensibility, length, and entropy distributions. Similarly, PLOS and eLife lay summaries have similar comprehensibility distributions but differ in their readability and length.

Model	Abstract						Lay summary - PLOS						Lay summary - eLife					
	R-1	R-2	R-L	BS	DCRS	FKGL	R-1	R-2	R-L	BS	DCRS	FKGL	R-1	R-2	R-L	BS	DCRS	FKGL
Heuristic Lead-3	23.86	5.66	21.48	81.17	12.66	14.82	27.41	6.87	24.61	83.36	12.66	15.08	19.41	4.06	18.02	81.65	12.65	13.30
Lead-K	35.69	9.07	32.70	82.86	11.69	14.49	38.28	34.8	33.72	83.72	11.88	14.95	37.27	7.53	35.18	82.05	10.58	11.89
Oracle	60.08	27.48	55.95	87.35	11.12	15.15	57.82	23.92	53.37	87.13	11.20	15.28	48.92	13.42	46.30	82.94	10.51	13.18
Unsup. TextRank	40.26	11.53	36.02	83.83	11.78	20.08	37.55	8.50	33.28	83.43	11.87	20.27	33.88	5.79	31.55	81.16	11.30	18.98
LexRank	38.22	13.06	35.42	83.85	9.70	14.23	31.20	9.09	28.72	82.97	9.70	14.59	32.25	5.73	30.45	80.67	9.68	13.32
HipoRank	36.95	10.19	33.89	83.22	12.15	14.46	37.67	9.22	34.28	83.68	12.15	14.69	31.50	5.17	29.68	80.88	12.13	12.13
Supervised BART	43.34	13.14	39.80	85.48	11.33	14.40	43.52	12.09	39.67	85.70	11.29	14.54	31.17	6.74	29.20	83.55	11.15	13.87
BART$_{Scaffold}$	43.13	12.87	39.66	85.33	11.10	14.14	43.73	12.22	39.92	85.67	11.30	14.58	43.01	10.82	40.54	84.88	9.68	11.85
GPT3.5-zs	28.69	6.52	15.04	82.76	11.70	14.32	42.74	12.70	22.28	86.32	10.40	13.19	33.72	8.45	16.95	84.36	10.36	13.03
GPT3.5-mdc	-	-	-	-	-	-	44.41	14.16	41.12	86.55	10.36	13.32	37.97	9.39	35.57	84.22	10.78	13.70
ATLAS	45.87	14.08	42.32	85.54	10.96	14.21	44.44	12.33	40.60	85.70	11.22	14.58	46.80	12.57	44.14	85.20	8.95	10.87
ATLAS$_{Oracle}$	46.11	14.07	42.51	85.69	10.99	14.13	44.97	12.49	41.02	85.82	11.21	14.48	46.61	12.29	43.95	85.11	9.18	11.39

Table 2: Summarization performance on the PLOS and eLife test sets (abstracts combined). R = ROUGE F1 (↑), BS = BERTScore (↑), DCRS = Dale-Chall Readability Score (↓), FKGL = Flesh-Kincaid Grade Level (↓). For supervised models, we highlight the best score obtained for each metric in **bold** and underline second best.

and eLife. We include the results for ATLAS under two conditions: 1) one utilizing the average value for each attribute observed in the training data for each summary type (ATLAS); and 2) one using true attribute values obtained from gold standard summaries (ATLAS$_{Oracle}$), where ATLAS$_{Oracle}$ is intended to provide an upper bound of the obtainable performance using our control attributes.

For all metrics, it is evident from Table 2 that ATLAS exceeds the performance of all baseline approaches for both eLife lay summaries and abstracts, demonstrating a strong ability to control the technicality of generated text whilst producing high-quality summaries. Interestingly, although the GPT3.5-mdc baseline achieves a slightly stronger all-round performance for PLOS lay summaries, it fails to maintain this for the more "lay" summaries of eLife where ATLAS achieves significantly better performance, indicating that our control attributes can effectively capture these differences.

In all cases, ATLAS also achieves scores that are comparable to (and sometimes exceeding) that of ATLAS$_{Oracle}$, suggesting that the use of the most frequently observed bin value for control attributes is effective for producing the appropriate characteristics for each summary type.

Ablation study. To assess the contribution of each attribute to model performance, we conduct an ablation study, evaluating ATLAS$_{Oracle}$ under different configurations.[4] Table 3 reports the results of this study for abstracts and lay summaries on the combined test sets of PLOS and eLife.

The table shows that the removal of control attributes has a significant detrimental effect on performance. Additionally, when only a single attribute is included, the length-based control has

Model	Lay summary				Abstract			
	R-1	R-2	R-L	DCRS	R-1	R-2	R-L	DCRS
BART	41.68	11.29	38.12	11.27	43.34	13.14	39.80	11.33
+R	43.34	12.03	39.75	10.91	43.49	13.23	39.95	11.12
+BG	42.52	11.71	39.01	11.01	43.74	13.65	40.35	10.98
+CWE	41.58	11.21	38.04	11.28	44.23	13.48	40.56	11.35
+L	44.22	12.21	40.55	10.81	44.83	13.75	41.31	11.03
+L+BG	44.66	12.36	40.96	10.99	45.67	13.78	42.02	11.17
+L+R	44.52	12.10	40.73	10.92	45.54	13.64	41.78	11.21
+L+CWE	44.72	12.41	41.04	10.88	45.87	13.99	42.32	10.10
+L+R+BG	44.82	12.41	41.10	10.97	45.94	14.07	42.32	11.10
+L+R+CWE	44.83	12.39	41.05	10.90	45.60	13.63	41.84	11.21
+L+BG+CWE	45.01	12.56	41.38	10.88	46.04	14.16	42.44	11.06
ATLAS$_{Oracle}$	45.22	12.47	41.45	10.91	46.11	14.07	42.51	10.99

Table 3: Ablation study on the ROUGE-based performance of ATLAS under different configurations using true attribute values. "+" denotes aspect addition. L = Length, R = Readability, CWE = Content Word Entropy, BG = Background information.

the highest ROUGE scores, particularly for lay summaries. This is to be expected, as lay summaries are known to differ significantly in length between PLOS (avg. 175.6 words) and eLife (avg. 347.6 words). When employing attributes in combination, we can see that the addition of content word entropy control and the subsequent addition of background information control have the greatest benefit to performance for ATLAS with 2 and 3 attributes, respectively. Interestingly, no attribute emerges clearly as the least effective as, although readability score control is the only one not included in the 3 attribute model, its inclusion in the single attribute model has clear benefits for lay summary performance. This provides further evidence that, in combination, our control attributes are able to capture the differences between summary types and effectuate them during generation.

Human evaluation. To provide a comprehensive assessment of the summaries generated, we conducted a human evaluation involving our proposed model ATLAS and the strongest baseline model

---

[4]We use ATLAS$_{Oracle}$ as the subject of this experiment rather than ATLAS to get a true reflection of each attribute's influence, rather than an approximation.

Criteria	eLife		PLOS	
	BART	ATLAS	BART	ATLAS
Comprehensiveness	2.30	2.65	2.00	2.55
Layness	2.60	3.05	2.10	2.45
Factuality	2.20	2.85	2.05	2.40

Table 4: Human evaluation on eLife and PLOS. Mean evaluator ratings (1-5) obtained by BART and ATLAS outputs for each metric.

	Model	FKGL	CLI	DCRS
PLOS	$ATLAS_{technical}$	15.11	14.21	11.64
	$ATLAS_{lay}$	13.22	13.97	11.22
eLife	$ATLAS_{technical}$	14.77	14.02	11.32
	$ATLAS_{lay}$	10.89	11.45	9.17

Table 5: Readability metrics for two versions of ATLAS with highly lay and technical attribute values.

(BART) using two experts.[5] Specifically, adopting a similar setting to the original that of Goldsack et al. (2022), we take a random sample of 10 articles from the test split of each dataset. Alongside each model-generated lay summary, judges are presented with both the abstract and reference lay summary of the given article. We choose not to provide judges with the full article text in an effort to minimise the complexity of the evaluation and the cognitive burden placed upon them. Using 1-5 Likert scale, the judges are asked to rate the model output based on three criteria: (1) *Comprehensiveness*: to what extent does the model output contain the information that might be necessary for a non-expert to understand the high-level topic of the article and the significance of the research; (2) *Layness*: to what extent is the content of the model output comprehensible (or readable) to a non-expert, in terms of both structure and language; (3) *Factuality*: to what extent is the model generated lay summary factually consistent with the two other provided summaries (i.e. abstract and reference lay summary).[6]

Table 4 presents the average ratings from our manual evaluation. We calculate the Cohan Kappa scores to measure inter-rater reliability, where we obtain values of 0.50 and 0.57 for eLife and PLOS, attesting to the reliability of our evaluation. The overall results suggest that our proposed method performs better than the BART baseline in terms of all three criteria on both datasets, attesting to their quality. In terms of layness, the higher layness scores observed in the eLife dataset compared to the PLOS dataset align with the previous analysis for the two datasets from (Goldsack et al., 2022). Moreover, compared to baseline, it is worth noting that our model outputs are judged to produce much more factually correct outputs on both datasets, suggesting our method generates fewer hallucinations.

Controllability analysis. To assess the extent to which our control attributes enable controllability over the overall layness of the text, we conduct a further analysis using two additional versions of ATLAS with highly lay or technical values. Specifically, we create $ATLAS_{lay}$ and $ATLAS_{technical}$ by selecting the lowest and highest attribute bins, respectively, for which there are at least 100 observations in the training data (for all attributes other than length which is kept constant).

We examine how these extreme attributes manifest themselves in generated summaries by calculating the average readability values obtained by the generated summaries for both datasets. We present the results of the analysis in Table 5, which show a significant divergence in the readability values obtained by each model on both datasets. Interestingly, this divergence is substantially wider for summaries generated on eLife, the dataset which is identified by Goldsack et al. (2022) as containing lay summaries that are more "lay" than those of PLOS, suggesting that exposure to more extreme values whilst training on this dataset may enable even greater controllability at inference time.[7]

5 Conclusion

In this paper, we introduce ATLAS, a model for controllable lay summarisation that employs controllable attribute tokens to influence various properties of the generated summary, enabling it to cater to users of different levels of expertise. Using combined datasets for biomedical lay summarisation we perform multiple experiments whereby we confirm the ability of our selected control attributes to discriminate between summary types, demonstrate their effectiveness for controllable lay summarisation, and further investigate their ability to effectuate desired differences during generation.

---

[5]Both judges have experience in scientific research and hold at least a bachelor's degree.

[6]For example, for the "Layness" criteria, a score of 5 is equal to "highly lay" and a score of 1, "highly technical".

[7]Examples of summaries generated by these models are included in the Appendices.

Limitations

Although our results demonstrate that our selected control attributes are able to effectively capture the characteristics between summary types, it is highly likely that there are additional attributes that we have not explored that could benefit performance for controllable lay summarisation. We plan to explore this in future work, in addition to experimenting with more complex methods for enabling controllability.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciB-ERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Nachshon Cohen, Oren Kalinsky, Yftah Ziser, and Alessandro Moschitti. 2021. Wikisum: Coherent summarization dataset for efficient human-evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 212–219. Association for Computational Linguistics.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. Discourse-Aware unsupervised summarization for long scientific documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1089–1102, Online. Association for Computational Linguistics.

G. Erkan and D. R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Tomas Goldsack, Zheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023a. 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, 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.

Tomas Goldsack, Zhihao Zhang, Chen Tang, Carolina Scarton, and Chenghua Lin. 2023b. Enhancing biomedical lay summarisation with external knowledge graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8016–8032, Singapore. Association for Computational Linguistics.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated Lay Language Summarization of Biomedical Scientific Reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):160–168.

Stuart R F King, Emma Pewsey, and Sarah Shailes. 2017. Plain-language Summaries of Research: An inside guide to eLife digests. *eLife*, 6:e25410.

Xiangzhe Kong, Jialiang Huang, Ziquan Tung, Jian Guan, and Minlie Huang. 2021. Stylized story generation with style-guided planning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2430–2436, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Zheng 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.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Chen Tang, Tyler Loakman, and Chenghua Lin. 2024. A cross-attention augmented model for event-triggered context-aware story generation. *Computer Speech & Language*, page 101662.

Oisín Turbitt, Robert Bevan, and Mouhamad Aboshokor. 2023. MDC at BioLaySumm task 1: Evaluating GPT models for biomedical lay summarization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 611–619, Toronto, Canada. Association for Computational Linguistics.

Liqiang Xiao, Lu Wang, Hao He, and Yaohui Jin. 2020. Modeling content importance for summarization with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3606–3611, Online. Association for Computational Linguistics.

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

A Appendix

ChatGPT Baseline Prompts The prompts provided to ChatGPT for each summary type are given in Table 6. To ensure a fair comparison, we control the length of the GPT baselines using the generation arguments, (e.g., max_new_tokens). Note that we differentiate between the lay summary types (namely, PLOS and eLife) based on distinctions made by Goldsack et al. (2022), who recognise PLOS' summaries as the less "lay" of the two, making them better suited to an audience with some technical knowledge.

Summary Type	Prompt
Abstract	Summarize the following article for an expert audience that is familiar with the technical aspects of the content
PLOS lay summary	Summarize the following article for a non-expert audience that has some familiarity with the technical aspects of the content
eLife lay summary	Summarize the following article for a non-expert audience that has no familiarity with the technical aspects of the content

Table 6: Prompts used for the GPT3.5-zs baseline for each summary type.

---

eLife

ATLAS$_{technical}$

The effects of muscle fatigue on motor learning under fatigue are poorly understood. Here, we investigated the effect of fatigue on learning under a sequential pinch force task. Irrespective of whether the observed fatigue effects are domain-specific or present in another task that is cognitive demanding but requires minimal force control, we found that participants had impaired skill learning in both the fatigued and unfatigued effector. We replicated the findings of experiment 1 and found that disruption of rTMS to the motor cortex (Cantarero et al ., 2013a) alleviated the adverse effects of fatigue. Cortical excitability was similar to that observed in the untrained effector, but not in the unfatigued. Altogether, our findings suggest that motor fatigue has a domain-dependent lasting effect on skill learning. Future studies should focus on understanding the role of motor cortex excitability in the acquisition of motor skills under fatigue, as well as the potential role for maladaptive memory formation under fatigued conditions. Cortical and motor cortices should be included in training and rehabilitation regimens geared to improve motor skill acquisition.

ATLAS$_{lay}$

Muscle fatigue is a neuromuscular phenomenon that can impair performance over time. People who experience fatigue tend to be less able to learn a new motor skill than people who experience no fatigue. However, it is not clear how fatigue affects the ability of people to learn new motor skills . One way to study the effects of fatigue is to study how people learn a motor skill under fatigue conditions. One of the main challenges in studying motor learning under fatigue is the so-termed "performance-learning" distinction In this study, participants were asked to practice a motor task over two days and then had to catch up to the skill performance level of the non-fatigued group. Unexpectedly , participants who were only fatigued at the end of the training were less likely to learn the motor skill. This suggests that fatigue has a domain-specific lasting effect on the learning of a skill. ernas et al. now show that people who are unable to recover the motor task under fatigue are more likely to be unable to learn their motor skill when they are not fatigued. The experiments show that when people are trained to perform the task, their ability to recover from fatigue is severely impaired. This effect is due to a change in the strength of the motor cortex, a region of the brain that is involved in learning and memory.

Figure 2: An case study from the eLife test set comparing summaries generated under highly lay and technical attribute values (with the length attribute being kept constant).

PLOS

<u>ATLAS$_{technical}$</u>

In this paper, we explore the conditions under which associations between antigenic, metabolic and virulence properties of strains within pneumococcal populations and predict how these may shift under vaccination. In this work , we use a conceptual framework to investigate the dynamics of associations between serotype, serotype and serotype-specific immunity in pneumococcus populations. We find that antigenic type (AT) is the principal determinant of non-capsular virulence factors (VF) , whereas MT is the major determinant. AT and MT are highly non-random; MT and AT are co-evolved and co-expressed. ET and CT are also found to be highly correlated, suggesting that they have synergistically adapted to a particular metabolic niche. IT and LD are found to have similar patterns of linkage disequilibrium (LD) than randomly selected genes not associated with metabolic/transport processes; AT is associated with a higher frequency of LD LD than MT LD; CT LD=0.013). CT is the first mathematical model to explain the non-overlapping association between serotypic and serotypes. TCT BC LD is a useful tool for predicting the potential impact of vaccination on the prevalence of serotypes associated with non-vaccine serotypes and for predicting how they may change under vaccination and vaccine serotype replacement.

<u>ATLAS$_{lay}$</u>

Pneumococcal populations are highly diverse in non-antigenic genes and are commonly classified into sequence types (ST) by Multi Locus Sequence Typing (MLST) of seven metabolic housekeeping genes. STs have been documented to occur regularly throughout the past 7 decades, yet many studies (eg) show an intriguing pattern of largely non-overlapping associations between serotype and ST. It has been noted that many STs that were previously associated with vaccine serotypes now occur in association with non-vaccine serotypes. It has been proposed that a combination of immune-mediated interference between identical antigenic types and direct competition between identical metabolic types can generate non-overlapping association between antigenic and STs in populations of the bacterial pathogen Neisseria meningitidis . In this paper, we explore whether pneumococcal population structure, can be explained within a similar conceptual framework. in which pathogen strains are profiled by antigenic type, AT, metabolic type (MT) and additional non-capsular virulence factors (VF).

Figure 3: An case study from the eLife test set comparing summaries generated under highly lay and technical attribute values (with the length attribute being kept constant).