Overview of the BioLaySumm 2023 Shared Task on Lay Summarization of Biomedical Research Articles

Tomas Goldsack[1], Zheheng Luo[2], Qianqian Xie[2],
Carolina Scarton[1], Matthew Shardlow[3], Sophia Ananiadou[2], Chenghua Lin[1]
[1]University of Sheffield, [2]University of Manchester, [3]Manchester Metropolitan University
{tgoldsack1, c.lin, c.scarton}@sheffield.ac.uk
{zheheng.luo, qianqian.xie, sophia.ananiadou}@manchester.ac.uk
m.shardlow@mmu.ac.uk

Abstract

This paper presents the results of the shared task on Lay Summarisation of Biomedical Research Articles (BioLaySumm), hosted at the BioNLP Workshop at ACL 2023. The goal of this shared task is to develop abstractive summarisation models capable of generating "lay summaries" (i.e., summaries that are comprehensible to non-technical audiences) in both a controllable and non-controllable setting. There are two subtasks: 1) Lay Summarisation, where the goal is for participants to build models for lay summary generation only, given the full article text and the corresponding abstract as input; and 2) Readability-controlled Summarisation, where the goal is for participants to train models to generate both the technical abstract and the lay summary, given an article's main text as input. In addition to overall results, we report on the setup and insights from the BioLaySumm shared task, which attracted a total of 20 participating teams across both subtasks.

1 Introduction

Biomedical publications report upon the latest research concerning prominent health-related topics, ranging from common illnesses to global pandemics (Wang et al., 2020). Accordingly, the content of these publications is of interest to a wide variety of audiences, including researchers, medical professionals, journalists, and even members of the public. However, the highly technical and specialist language used within such articles typically makes it difficult for non-expert audiences to understand their contents. This results in useful knowledge and findings having limited accessibility to the general public (Guo et al., 2021; Goldsack et al., 2022; Luo et al., 2022b).

Abstractive summarisation models can be used to generate a concise summary of an article, capturing its most salient points using words and sentences that do not necessarily appear in the original text of the article. As such, these models have

the potential to make highly technical documents accessible to a much wider audience through the generation of "lay summaries" — more readable summaries consisting largely of background information and containing minimal technical terminology (Guo et al., 2021; Goldsack et al., 2022; Luo et al., 2022b).

The BioLaySumm shared task[1] focuses on the abstractive summarisation of biomedical articles whilst placing an emphasis on controllability and ensuring comprehensibility for non-expert audiences. Through this shared task, we aim to foster increased research interest in Lay Summarisation (in both controllable and non-controllable settings), enabling further progression for novel model development and high-quality dataset construction. In turn, we hope this will help to broaden the accessibility of technical texts to non-specialist audiences and to drive progress towards more usable and effective abstractive summarisation models for the biomedical domain with the ability to cater to audiences possessing different levels of expertise.

In this paper, we present the results of the first BioLaySumm shared task, hosted by the BioNLP Workshop at ACL 2023. We cover the task formulation (§2), datasets (§3), and evaluation procedure (§4), before providing a description of the participating systems, overall results, and notable insights (§5).

2 Task Description

The shared task is composed of two separate subtasks, focusing on 1) the generation of summaries more suitable for a lay audience (Lay Summarisation), and 2) the development of controllable summarisation models capable of catering to audiences with different levels of expertise (Readability-controlled Summarisation).

---

[1]https://biolaysumm.org

2.1 Subtask 1: Lay Summarisation

Given an article's abstract and main text as input, the goal is for participants to train a model (or models) to generate the lay summary. Two separate datasets, **PLOS** and **eLife** (derived from the eponymous biomedical journals), were provided for model training and will be used for evaluation (more details on datasets are given in §3). For the evaluation, we average submission performance across both datasets.

For this task, we allowed submissions to be generated from either two separate summarisation models (i.e., one trained on each dataset) or a single unified model (i.e., trained on both datasets). Participants were required to indicate which approach was taken for each submission, in addition to whether or not they made use of additional training data (i.e., data not provided specifically for the task).

2.2 Subtask 2: Readability-controlled Summarisation

Given the main text of an article as input, the goal is for participants to train a model (or models) to generate both the technical abstract and the lay summary. A single dataset, **PLOS**, is provided for training and evaluation. We allowed submissions to use multiple ensemble models but still generate technical summary and the lay summary from the same model, and also one single main model with different output layers to generate two different summary types. As with subtask 1, participants are required to indicate whether or not they made use of additional training data for each submission. For the evaluation, we average submission performance across both summary types.

3 Datasets

The datasets used within each subtask are based on the previous works of Goldsack et al. (2022) and Luo et al. (2022b), and are derived from two different biomedical publications: **Public Library of Science (PLOS)** and **eLife**. Each dataset consists of research articles, their technical abstracts, and their expert-written lay summaries. As detailed in §2, each form of summary within these datasets (i.e., abstract and lay summary) has a different utility in each subtask. The lay summaries of each dataset also exhibit numerous notable differences in their characteristics, with eLife's lay summaries being longer, more abstractive, and more readable

Dataset	Subtask	# Train	# Val	# Test
eLife	1	4,346	241	142
PLOS	1, 2	24,773	1,376	142*

Table 1: Data split sizes for each dataset. * denotes that this split is different for each subtask.

than those of PLOS. Furthermore, for PLOS, lay summaries are author-written, and articles are derived from 5 peer-reviewed journals covering Biology, Computational Biology, Genetics, Pathogens, and Neglected Tropical Diseases. For eLife, lay summaries are written by expert editors (in correspondence with authors), and articles are derived from the peer-reviewed eLife journal that covers all areas of the life sciences and medicine. For more detailed analysis of dataset content, please refer to Goldsack et al. (2022).

Table 1 summarises the data split information for both datasets. Note that the training and validation sets used for both datasets are equal to those published in Goldsack et al. (2022). Furthermore, that the training and validation splits of PLOS are the same for both subtasks.

Alternatively, we collect new test splits for both PLOS and eLife data using more recently published articles from each respective journal. The test data for Subtask 1 is composed of 142 PLOS articles and 142 eLife articles. The test data for Subtask 2 is composed of 142 PLOS articles (however, these are different from those used in Subtask 1).

In utilising these datasets for our task, we hope to enable the training of abstractive summarisation models that are capable of generating lay summaries for unseen articles covering a wide range of biomedical topics, enabling the significance of new, important publications to be effectively communicated to non-expert audiences.

4 Evaluation

For both subtasks, we evaluate summary quality according to three criteria - *Relevance*, *Readability*, and *Factuality* - where each criterion is composed of one or more automatic metrics:

- *Relevance*: ROUGE-1, 2, and L (Lin, 2004) and BERTScore (Zhang et al., 2020b).

- *Readability*: Flesch-Kincaid Grade Level (FKGL) and Dale-Chall Readability Score (DCRS).

- *Factuality*: BARTScore (Yuan et al., 2021), fine-tuned on our respective datasets (as has proven effective in recent work (Koh et al., 2022)).[2]

For Subtask 1, the scores calculated for each metric are the average of those calculated independently for the generated lay summaries of PLOS and eLife. The aim is to maximise the scores for Relevance and Factuality metrics and minimise scores for Readability metrics.

For Subtask 2, the scores presented for each metric are the average of those calculated independently for the generated abstracts and lay summaries. Notably, for Readability metrics in this subtask, we calculate the *absolute difference* between the scores of generated summary and target summary pairs (rather than simply using the scores obtained for generated summaries, as in subtask 1). The aim is to maximise the scores for Relevance and Factuality metrics and minimise the absolute difference scores calculated for Readability metrics.

Following the submission deadline for each subtask, an overall ranking is calculated based on the cumulative rank of the evaluation criteria, where a lower overall ranking equates to better overall performance). To produce a criterion ranking, we apply min-max normalisation to the scores of each metric, before averaging across metrics within each evaluation criterion.

5 Shared Task Submissions

For both subtasks, we include a baseline system based on BART-base (Lewis et al., 2020) in order to provide a simple, widely-used benchmark with which submission performance can be compared. For subtask 1, this baseline system is composed of two separate BART models, trained independently on the PLOS and eLife datasets. For subtask 2, the baseline system is a controllable BART model, trained to generate either the abstract or lay summary of an article based on the inclusion of control tokens prepended to the input document ([ABSTRACT] and [SUMMARY], respectively). Participating teams for each subtasks were allowed to make a maximum of 3 submissions in total.

5.1 Submissions to Subtask 1

5.1.1 Systems Overview

Subtask 1 attracted a total of 20 participating teams, between them making a total of 49 submissions. A brief explanation of the modelling approach taken by each team is given below:[3]

LHS712EE (Liu et al., 2023) The team employed a BART model for eLife and a Longformer Encoder-Decoder (LED) model (Beltagy et al., 2020) for PLOS, whilst also experimenting with optimising memory usage.

GRASUM (Rosati, 2023) Standing for **G**rounded, **R**eferenced, and **A**nnotated **SUM**maries, this team's method combines approaches from retrieval augmentation, offline RL, and controlled generation, using a LED model with 16k input limit as the base model. A "grounding" step enhances each document with content retrieved from scientific abstracts, Wikipedia, simple Wikipedia, and UMLS that is appended to input, in addition the bibliographic reference string of the source document (obtained from CrossRef). An additional "annotation" step annotates each source document with control tokens that indicate whether the corresponding summary achieves higher or lower than the median score for each of the task evaluation aspects (given in §4).

BITSpi This team's method involves fine-tuning two separate BART models on pre-processed versions of each dataset. Specifically, stopwords are removed from the input data, and abbreviations are substituted for their full forms using a medical dictionary.

APTSum (Poornash et al., 2023) A three-step approach is adopted by this team, leveraging the SimCLS contrasting learning framework (Liu and Liu, 2021). Specifically, they first perform content selection, identifying the Abstract and Introduction as best model input, before generating candidate summaries using BART, followed section-wise re-ranking using a RoBERTa-base model to capture section-based salience information.

LaSTUS-FBK This team used a multi-stage unified approach, first cleaning the data via reference

---

[2]A fine-tuned version of the FactCC (Kryscinski et al., 2020) metric was also originally included for Factuality evaluation. However, preliminary testing found that it did not provide a reliable indication of factual correctness for the task.

[3]Note that we were unable to get a response from every team describing their modelling approach, hence there are some teams missing from this section.

Team	#	+	Relevance				Readability		Factuality
			R-1	R-2	R-L	BERTs	FKGL	DCRS	BARTs
himil	2	×	**49.46**	15.68	45.91	85.85	13.17	10.14	-2.41
Path. Dynamics	2	×	49.38	15.93	45.97	85.93	13.10	10.12	-2.33
Marsfield_SDS	2	✓	49.33	16.24	**46.15**	86.10	12.55	9.84	-2.25
LHS712EE	2	×	49.27	15.75	45.84	86.57	13.31	10.22	-1.12
APTSumm	2	✓	48.32	14.91	45.41	84.30	12.22	9.00	-3.41
VBD-NLP	2	×	48.29	14.69	45.02	85.71	12.29	10.09	-1.74
MDC	1	×	48.22	15.53	44.85	**87.07**	12.94	10.21	-1.18
HanyangLab	2	×	48.18	14.18	44.20	85.83	12.83	10.50	-1.92
Arizona Sky	2	×	48.11	14.50	44.42	85.75	12.36	10.12	-1.80
ViNLPSum	-	-	47.97	15.46	44.85	85.77	12.77	9.68	-2.01
GRASUM	1	✓	47.69	**16.76**	44.30	86.01	12.73	10.50	-2.33
IKM_Lab	1	×	47.44	14.45	44.31	85.66	11.84	9.84	-2.34
baseline	2	×	46.96	14.45	43.71	86.42	12.07	10.25	**-0.83**
NCUEE-NLP	2	×	45.87	13.41	41.84	85.47	12.94	10.48	-2.71
HUST-NLP	-	-	43.29	12.25	39.22	85.30	12.69	10.79	-1.91
IITR	1	×	42.81	11.17	32.15	85.14	**10.70**	**8.45**	-1.79
noobitA	-	-	42.75	11.50	39.45	85.77	12.72	10.41	-1.83
LaSTUS-FBK	1	×	39.98	10.46	36.53	84.79	15.63	11.25	-2.33
ISIKSumm	1	×	37.06	9.04	34.48	82.55	12.13	10.02	-3.92
BITSpi	2	×	37.03	10.77	33.66	84.97	11.77	10.97	-2.97
nippon	-	-	36.53	8.82	33.62	83.44	12.80	10.42	-2.28

Table 2: Subtask 1 leaderboard - all metrics. The **#** column denotes the number of models used - 1 (unified) or 2 (one for each dataset), and the **+** column denotes the use of additional training data. "-" indicates that the corresponding information was not provided. **R** = ROUGE F1, **BERTs** = BERTScore, **FKGL** = Flesch-Kincaid Grade Level, **DCRS** = Dale-Chall Readability Score, **BARTs** = BARTScore.

removal and acronym resolution. Extractive summarisation based on similarity-based sentence classification is then used to shorten the input before the resulting text is enhanced with the injection of complex concept definitions from Wikipedia. Finally, abstractive summarisation is performed using a fine-tuned BART model pre-trained on PubMed on a dataset-balanced sample of the training data (4K training instances from each dataset).

Marsfield_SDS (Sim et al., 2023) Using two fine-tuned FLAN-T5 models (one for each dataset) as the backbone of their experiments, this team experimented with different data augmentation strategies including the use of ChatGPT for paraphrasing existing lay summaries.

VBD-NLP (Phan et al., 2023) This team's method is based on the combined use of sequence-to-sequence model BioBART (Yuan et al., 2022) and FACTORSUM (Fonseca et al., 2022), a factorized energy-based model that aims to identify the most important input content, enabling more effective processing of long documents. Additional

experimentation with handling length as well as utilising other Pretrained Language Models (PLMs) was also carried out.

MDC (Turbitt and Bevan, 2023) This team focused on comparing the performance of general-purpose GPT models (e.g., ChatGPT) with in-domain GPT models (e.g., BioGPT (Luo et al., 2022a)). Additionally, they experimented with zero-shot and few-shot prompting, as well as fine-tuning different models.

Pathology Dynamics (Al-Hussaini et al., 2023) The team experimented with multiple different approaches based on BART and T5 models including methods of content selection, the use of efficient attention mechanisms (to better process long documents), and the zero-shot simplification of model outputs. Of those tested, the approach that achieved the best overall performance was BART-large, pre-trained on CNN-DM dataset, with inputs truncated to 1024 tokens.

Pos.	Team	Relevance	Readability	Factuality	Sum
1	**MDC**	3	10	3	**16**
2	**baseline**	9	8	**1**	18
3*	**Marsfield_SDS**	2	6	11	19
3*	**VBD-NLP**	8	7	4	19
5	**LHS712EE**	**1**	18	2	21
6	**ViNLPSum**	7	5	10	22
7	**IITR**	17	**1**	5	23
8	**Arizona Sky**	10	9	6	25
9	**Path. Dynamics**	4	11	14	29
10	**IKM_Lab**	11	3	16	30
11	**himil**	5	13	17	34
12*	**APTSumm**	13	2	20	35
12*	**GRASUM**	6	16	13	35
14	**noobitA**	16	13	7	36
15	**HanyangLab**	12	17	9	38
16	**HUST-NLP**	15	20	8	43
17	**ISIKSumm**	21	4	21	46
18	**nippon**	20	15	12	47
19	**NCUEE-NLP**	14	29	18	51
20	**BITSpi**	19	14	19	52
21	**LaSTUS-FBK**	19	21	15	54

Table 3: Subtask 1 leaderboard - criteria rankings.

IITR (Reddy et al., 2023)   Also using BART and T5 models trained on both dataset, this team experiment with different methods of content selection and ordering.

Arizona Sky   This team first truncate input documents, before using them to train two separate BART base models.

IKM_Lab (Wu et al., 2023)   This team experimented with the use of a LED model trained on both datasets, as well as the adoption of different formats for including additional article information, such as keywords and section headings, in the input.

NCUEE-NLP (Chen et al., 2023)   This team also made use of different models for each submission, including Primera (Xiao et al., 2022), a PEGASUS model (Zhang et al., 2020a) pretrained on PubMed, and a BART-large Longformer model.

himil   The team experimented with both BERT (Devlin et al., 2019) and Longformer-based models, trained individually on each dataset.

5.1.2   Results

Table 2 presents the performance of the submission selected to appear on the leaderboard by each team according to the defined task metrics and Table 3 presents the rankings of these submissions (both overall and according to each individual criteria) following the application of the evaluation process described in §4.

In general, we find that more teams opted for the use of two models (10 out of 20), one for each of the two provided datasets, rather than a single unified model trained on both datasets (6 out of 20). Furthermore, the use of additional training data (i.e., data not provided as part of this task) to directly fine-tune models was relatively rare, with only 3 confirmed instances. However, all participants decided to make use of pre-trained language models (PLMs) in their submissions. In terms of the specific models used, we find BART-based models (e.g., BART, Longformer Encoder-Decoder, etc.) to be a particularly popular choice amongst teams, being utilised by 11 out of 13 teams who provided detailed descriptions of their method. Finally, we observe that several teams also chose to experiment

with data preprocessing, implementing methods such as data cleaning, data annotation, and data augmentation with varying degrees of success.

We find that the best overall system (i.e., that which achieved the lowest summed ranking across the three evaluation criteria) is that of team MDC, whose best submission utilises a single ChatGPT-based model (text-davinci-003) coupled with few-shot prompting to generate the lay summaries of both datasets, based on only the abstracts. Although this system does not achieve the best performance in any individual criteria, it achieves a strong performance for both Relevance and Factuality (ranking 3rd for both) whilst maintaining an above-average Readability ranking (10th). The fact that the cumulative rank of this system is equal to 16 is evidence that no model is able to achieve universally strong performance across all criteria (relative to other submissions). However, the fact that the top-ranking submission is based on only few-shot in-context learning (i.e., without any fine-tuning on the provided training data) suggests that Large Language Models have the potential to offer significant benefits for Lay Summarisation.

Interestingly, this is the only submission to achieve a better cumulative rank than that of the BART baseline system (18th), which is shown to rank first for Factuality, and above average for the other two criteria (9th and 8th for Relevance and Readability, respectively). We originally suspected that a possible explanation for the baseline system's strong performance in terms of Factuality is a potential bias of BARTScore towards BART-based models. However, the leaderboard results do not seem to support this, with BART-based models being widely used and achieving a wide range of scores.

Two teams tied for third in terms of overall ranking, with both Marsfield_SDS and VBD-NLP achieving a cumulative rank of 19. Each of these teams adopted innovative and diverse strategies with their submissions. Marsfield_SDS focused largely on data augmentation including the use of ChatGPT for generating lay summary paraphrases, resulting in particularly strong performance in terms of Relevance (2nd). Alternatively, VBD-NLP experimented with the use of the factorised energy-based model FACTORSUM, achieving a good all-rough performance across all criteria. Finally, the 5th placed submission of team LHS712EE is also worthy of note, obtaining the

best rank for Relevance and 2nd best for Factuality.

5.2 Submissions to Subtask 2

5.2.1 Systems Overview

Three teams have made in total 7 attempts for Subtask 2. A brief description of their respecitve approaches are as following:

LHS712EE (Liu et al., 2023) The team carried on with the LED (Beltagy et al., 2020) model trained on the PLOS dataset from Subtask 1 to test the generalizability of their approach in generating lay summaries coupled with a pre-trained LED model for abstractive summaries. They later retrained the model using the abstract section of the dataset to improve performance in generating technical abstracts.

Pathology Dynamics (Al-Hussaini et al., 2023) As the abstract with the most salient information is no longer present in the input, to tackle the long context input, the team trained a base LSG model (Condevaux and Harispe, 2022) and truncated each article to the first 4096 tokens for generating both abstracts and lay summaries. The model was then trained on a merged dataset that uses each article twice, with one output having the lay summaries and the other having the abstract. They also reported using simplification procedures such as MUSS (Martin et al., 2022) to enhance the lay summary or other instruction-following models such as T5 with different prefix for summarisation.

NCUEE-NLP (Chen et al., 2023) This team made use of different models for each submission, including Primera, a PEGASUS model pre-trained on PubMed, and a BART-large Longformer model.

5.2.2 Results

In Table 4, the performance of the submissions to Subtask 2 is shown on the leaderboard by each team according to the defined task metrics. Table 5 presents the overall and by individual metric rankings of these submissions following the application of the evaluation process described in §4. Due to the overall ranking scheme and the limited number of participants, we have all three teams ranked first while demonstrating advantages and disadvantages in different aspects.

All three teams utilise augmented transformers that can take longer input context, which significantly boosts the performance of Rouge score

Team	#	+	Relevance				Readability		Factuality
			R-1	R-2	R-L	BERTs	FKGL	DCRS	BARTs
LHS712EE	2	×	44.17	12.99	40.53	**85.49**	2.263	0.9364	-1.1403
NCUEE-NLP	1	×	**45.14**	**14.02**	**41.23**	85.45	**2.047**	0.9340	-2.1102
Pathology Dynamics	1	×	45.11	13.82	41.00	85.32	2.106	**0.8232**	-1.5682
<u>baseline</u>	1	×	40.88	11.63	36.86	**85.49**	2.396	0.9312	**-0.9783**

Table 4: Subtask 2 leaderboard-all metrics. The **#** column denotes the number of models used - 1 (unified) or 2 (one for each dataset), and the **+** column denotes the use of additional training data. **R** = ROUGE F1, **BERTs** = BERTScore, **FKGL** = Flesch-Kincaid Grade Level, **DCRS** = Dale-Chall Readability Score, **BARTs** = BARTScore.

Pos.	Team	Relevance	Readability	Factuality	Sum
1*	**NCUEE-NLP**	1	2	4	7
1*	**Pathology Dynamics**	3	1	3	7
1*	**LHS712EE**	2	3	2	7
4	<u>**baseline**</u>	4	4	1	9

Table 5: Subtask 2 leaderboard - criteria rankings.

while also achieving smaller readability differences. We assume this is because longer input enables the models to see more lexicons that can be used to build summaries, resulting in a better chance to overlap with the reference summaries. However, these improvements do not necessarily promise higher results on LM-based metrics such as BERTScore and BARTScore on which the baseline method prevails.

It is worth noting that Team Pathology Dynamics used summaries generated from a LSG model simultaneously trained on both plain language as well as technical references and get output as a hybrid of the lay summaries and abstracts. Their methods obtains the highest readability and joint overall highest scores, suggesting the limitation of the readability metrics used for evaluation. In addition, they reported that neither simplification model nor small-scale instruction-following models succeed to improve performance in this task.

In conclusion, none of the participating team secured a sweeping superiority across the three evaluated aspects, highlighting the challenge in readability-controlled summarisation on relatively small-scaled language models. Given that LLMs (Large Language Models) better align with human instructions (Ouyang et al., 2022), we expect future work to investigate their capabilities in the task.

6 Conclusion

The first BioLaySumm shared task was hosted at the BioNLP Workshop @ ACL2023 and consisted of two subtasks focusing on Lay Summari-

sation and Readability-controlled Summarisation, respectively. The task attracted a total of 20 teams, between them making 56 individual submissions across both subtasks. Submissions were evaluated according to three general criteria - Relevance, Readability, and Factuality — with each criteria consisting of one or more automatic metrics.

The results of both subtasks show that achieving strong performance for all three criteria (relative to other submissions) was particularly rare, attesting to the challenging nature of generating lay summaries for research articles in both controlled and non-controlled settings. Furthermore, when also taking into account the relatively strong performance of the BART baseline models (in particular for the Factuality component of our evaluation), this suggests that further research effort is required to develop truly usable models that can be reliably deployed in real-world settings.

However, as demonstrated by highly-ranked teams MDS and Marsfield_SDS (who obtain first and joint third-ranking submissions for subtask 1, respectively), recent developments in the abilities of both general-purpose and in-domain LLMs have the potential to offer significant benefits for the automatic generation lay summaries. As such, we expect that utilising such models for summary generation, data augmentation, and evaluation to be promising future directions for Lay Summarisation.

References

Irfan Al-Hussaini, Austin X. Wu, and Cassie Mitchell. 2023. Pathology dynamics at biolaysumm: the trade-off between readability, relevance, and factuality in lay summarization. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

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

Chao-Yi Chen, Jen-Hao Yang, and Lung-Hao Lee. 2023. Ncuee-nlp at biolaysumm task 2: Readability-controlled summarization of biomedical articles using the primera models. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Charles Condevaux and Sébastien Harispe. 2022. Lsg attention: Extrapolation of pretrained transformers to long sequences. *ArXiv*, abs/2210.15497.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcio Fonseca, Yftah Ziser, and Shay B. Cohen. 2022. Factorizing content and budget decisions in abstractive summarization of long documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6341–6364, Abu Dhabi, United Arab Emirates. 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.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022. How far are we from robust long abstractive summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2682–2698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, 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*, pages 7871–7880, Online. 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.

Quancheng Liu, Xiheng Ren, and V. G. Vinod Vydiswaran. 2023. Lhs712ee at biolaysumm 2023: Using bart and led to summarize biomedical research articles. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022a. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6). Bbac409.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022b. 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.

Louis Martin, Angela Fan, Eric Villemonte de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. Muss: Multilingual unsupervised sentence simplification by mining paraphrases. In *International Conference on Language Resources and Evaluation*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to

follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Phuc Phan, Tri Huu Tran, and Hai-Long Trieu. 2023. Vbd-nlp at biolaysumm task 1: Explicit and implicit key information selection for lay summarization on biomedical long documents. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

A.S. Poornash, Atharva Deshmukh, Archit Sharma, and Sriparna Saha. 2023. Aptsumm at biolaysumm task 1: Biomedical breakdown, improving readability by relevancy based selection. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Venkat praneeth Reddy, Pinnapu Reddy, Harshavardhan Reddy, Karanam Sai Sumedh, and Raksha Sharma. 2023. Iitr at biolaysumm task 1:lay summarization of biomedical articles using transformers. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Domenic Rosati. 2023. Grasum at biolaysumm task 1: Background knowledge grounding for readable, relevant, and factual biomedical lay summaries. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Mong Yuan Sim, Maciej Rybinski, and Sarvnaz Karimi. 2023. Csiro data61 team at biolaysumm task 1: Lay summarisation of biomedical research articles using generative models. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Oisín Turbitt and Robert Bevan. 2023. Mdc at biolaysumm task 1: Evaluating gpt models for biomedical lay summarization. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Yu-Hsuan Wu, Ying-Jia Lin, and Hung-Yu Kao. 2023. Ikm_lab at biolaysumm task 1: Longformer-based

prompt tuning for biomedical lay summary generation. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

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

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Appendix

Team	Relevance				Readability		Factuality
	R-1	R-2	R-L	BERTs	FKGL	DCRS	BARTs
himil	**1**	0.864	0.983	0.730	0.491	0.601	0.488
Path. Dynamics	0.994	0.895	0.987	0.748	0.476	0.596	0.515
Marsfield_SDS	0.990	0.935	**1**	0.785	0.364	0.496	0.541
LHS712EE	0.985	0.873	0.978	0.889	0.521	0.631	0.906
APTSumm	0.912	0.767	0.947	0.387	0.295	0.192	0.167
VBD-NLP	0.909	0.739	0.919	0.699	0.310	0.583	0.707
MDC	0.904	0.845	0.907	**1**	0.443	0.625	0.888
HanyangLab	0.901	0.675	0.861	0.726	0.422	0.730	0.649
Arizona Sky	0.896	0.715	0.876	0.708	0.324	0.592	0.685
ViNLPSum	0.885	0.826	0.907	0.712	0.408	0.438	0.618
GRASUM	0.863	**1**	0.868	0.765	0.401	0.732	0.516
IKM_Lab	0.844	0.775	0.866	0.688	0.217	0.495	0.512
BART Baseline	<u>0.807</u>	<u>0.709</u>	<u>0.826</u>	<u>0.856</u>	<u>0.264</u>	<u>0.641</u>	**<u>1</u>**
NCUEE-NLP	0.722	0.578	0.692	0.646	0.445	0.724	0.393
HUST-NLP	0.523	0.432	0.505	0.608	0.391	0.833	0.650
IITR	0.486	0.296	0	0.573	**0**	**0**	0.690
noobitA	0.481	0.338	0.521	0.712	0.399	0.698	0.678
LaSTUS-FBK	0.267	0.207	0.313	0.496	1	1	0.514
ISIKSumm	0.041	0.028	0.166	0	0.277	0.560	0
BITSpi	0.039	0.246	0.108	0.535	0.203	0.897	0.309
nippon	0	0	0.105	0.197	0.416	0.701	0.533

Table 6: Subtask 1 leaderboard - metric values normalised using min-max normalisation, so values range from 0-1.

Team	Relevance				Readability		Factuality
	R-1	R-2	R-L	BERTs	FKGL	DCRS	BARTs
NCUEE-NLP	**1**	**1**	**1**	0.764	**0**	0.978	0
Pathology Dynamics	0.992	0.916	0.947	0	0.169	**0**	0.478
LHS712EE	0.772	0.569	0.839	**1**	0.619	1	0.856
Baseline	0	0	0	**1**	1	0.954	**1**

Table 7: Subtask 2 leaderboard - metric values normalised using min-max normalisation, so values range from 0-1.