Overview of the BioLaySumm 2024 Shared Task on the Lay Summarization of Biomedical Research Articles

Tomas Goldsack[1], **Carolina Scarton**[1], **Matthew Shardlow**[3], **Chenghua Lin**[1,2]
[1]University of Sheffield, [2]University of Manchester, [3]Manchester Metropolitan University
{tgoldsack1, c.lin, c.scarton}@sheffield.ac.uk
m.shardlow@mmu.ac.uk

Abstract

This paper presents the setup and results of the second edition of the BioLaySumm shared task on the Lay Summarisation of Biomedical Research Articles, hosted at the BioNLP Workshop at ACL 2024. In this task edition, we aim to build on the first edition's success by further increasing research interest in this important task and encouraging participants to explore novel approaches that will help advance the state-of-the-art. Encouragingly, we found research interest in the task to be high, with this edition of the task attracting a total of 53 participating teams, a significant increase in engagement from the previous edition. Overall, our results show that a broad range of innovative approaches were adopted by task participants, with a predictable shift towards the use of Large Language Models (LLMs).

1 Introduction

Lay Summarisation describes the task of generating a summary of a technical or specialist text that is suitable for a non-expert audience. To achieve this goal, a good lay summary will typically focus on explaining the relevant background information alongside the significance and findings of an article, while avoiding extensive use of jargon or technical language. As such, lay summaries offer significant benefits in broadening access to technical articles that are of interest to a broad range of audiences.

Biomedical research publications, containing the latest research on prominent health-related topics, represent a perfect example of such texts. Not only are the contents of these articles relevant to other researchers working in the same domain, but often they can also be of interest to researchers in related domains, medical practitioners, and even members of the public (e.g., those affected by an illness/disease being studied). In this scenario, the lay summary is an essential tool in allowing these secondary audiences, who don't possess the expertise required to interpret the full article, to understand its key findings and relevance to their information needs.

Although Lay Summaries are required or encouraged by some biomedical publications, they are not universal, leaving a significant amount of inaccessible to lay audiences. Furthermore, the burden of writing these summaries is often placed upon the article authors, who are not always adept at effectively communicating their work to a non-technical audience. As such, automatic lay summary generation has significant potential to help both authors and non-expert readers by improving the dissemination of important research.

The BioLaySumm shared task[1] focuses on improving the automatic lay summarization of biomedical research. Through this shared task, we aim to encourage the development of novel approaches and increase interest in the research of automatic techniques for disseminating scientific research to broad audiences. In this paper, we present the results of the second edition of the BioLaySumm shared task, hosted by the BioNLP Workshop at ACL 2024.

In what remains of the paper, we address the task formulation (§2), datasets (§3), and evaluation procedure (§4), before providing a description of overall results and notable insights (§5), and finall the participating systems (§6).

2 Task Description

As part of the task, participants must develop systems capable of generating a lay summary of biomedical research, given the article's text as input. Our competition was hosted using the CodaBench platform (Xu et al., 2022).

As with the previous edition of the task, two separate datasets, **PLOS** and **eLife** are used. Participants were provided with both training and valida-

[1] https://biolaysumm.org

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 122–131

tion sets, complete with reference lay summaries, alongside a blind test set. Final system performance is determined by the performance of participants' systems on the blind test set, which could be obtained by submitting their predicted lay summaries to our CodaBench competition, where they were automatically evaluated. More information regarding our datasets and evaluation protocol is provided in subsequent sections (§3 and §4, respectively).

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). Participants were also allowed to compete as part of teams, where each team was permitted to make a maximum of 15 test set submissions to the CodaBench platform.[2] However, teams were required to select only one of their submissions to appear on the final task leaderboard.

Because of its strong performance in the previous edition of the task, we also choose to keep the same baseline system. Specifically, this baseline system consists of two separate BART-base models (Lewis et al., 2020), trained independently on the PLOS and eLife datasets.

3 Datasets

The datasets used for the task 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 biomedical research articles paired with expert-written lay summaries.

As described in Goldsack et al. (2022), 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 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, covering all areas of the life sciences and medicine. For a more detailed analysis of dataset content, readers can refer to Goldsack et al. (2022).

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.

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).

As with the previous task edition, we collect new test splits for both PLOS and eLife data using more recently published articles from each respective journal. This test data consists of 142 PLOS articles and 142 eLife articles.

In utilizing 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), Dale-Chall Readability Score (DCRS), *Coleman-Laiu Index (CLI), and *LENS (Maddela et al., 2023).
- *Factuality*: *AlignScore (Zha et al., 2023) and *SummaC (Zha et al., 2023)

Where "*" indicates that the metric is newly introduced for this year's edition of the task. Specifically, the CLI and LENS metrics are introduced in order to enhance our evaluation of summary readability. Alternatively, AlignScore and SummaC are introduced to replace the fine-tuned BARTScore

[2] A significant increase on the limit of 3 submissions imposed in the previous edition of the task.

model used to assess factuality in the previous task edition, with the reason for this being that BARTScore was found to exhibit bias toward BART-based approaches.

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 maximize the scores for all metrics except for FKGL, DCRS, and CLI the Readability metrics. For these metrics, the aim is to minimize scores, as lower scores are indicative of greater readability.[3]

Following the submission deadline for each subtask, an overall ranking is calculated based on the average performance of submissions across all criteria. Specifically, we first apply min-max normalization to the scores of each metric (thus establishing a common value range), before averaging across metrics within each criterion to obtain criterion-level scores.[4] Note that, for metrics that we minimize (i.e., FKGL, DCRS, and CLI) we calculate 1 minus the mix-max normalized value. Finally, criterion-level scores are then averaged to obtain an overall score, by which submissions are then ranked.

5 Task Results

Table 2 presents the performance of the submission selected by each team to appear on the final leaderboard, according to the defined task metrics.

Overall, the final leaderboard of BioLaySumm 2024 contains a total of 53 teams, who made a combined total of over 200 submissions. This represents a 165% increase in participation over BioLaySumm 2023, which attracted a total of 20 teams across two subtasks. In this section, we summarize some of the key results and notable trends that were observed among participants.

Model selection trends We identify several trends amongst participants in terms of the models used for experiments.[5] Firstly, the use of Large Language Models was found to be particularly prevalent, with a total of 18 teams indicating that LLMs were used in some capacity. Compared to the 3 teams who used LLMs in BioLaySumm 2023, this represents a stark increase that is reflective of shifts in the broader research landscape of NLP. Within those teams using LLMs, biomedical-specific models such as BioGPT (Luo et al., 2022a) and BioMistral (Labrak et al., 2024) proved popular, with 7 teams indicating they used such models. Other LLMs used include GPT-4 (2), LLAMA (2), and Claude (1). There is evidence that LLMs were used for both summary generation and summary post-processing, with various settings (including fine-tuned, few-shot, and zero-shot) being adopted.

Outside of LLMs, the T5 (Raffel et al., 2019) model family proved the most popular alternative approach, with 13 teams making use of these models in their selected submissions. In particular, the FLAN-T5 (Chung et al., 2022a) model was found to be widely-used, being selected by 9 teams. Interestingly, only 3 teams were found to have used BART-based models, a significant drop from the previous BioLaySumm edition, where they were the most widely adopted approach. We find this shift in model selection to be an encouraging sign that participants are keen to explore novel methods for Lay Summarisation, in line with our overall task objectives.

Baseline comparison As shown in Table 2, 5 teams exceed the overall rank of the BART baseline system. This represents an an increase on the previous edition of the task, whereby only 1 team outperformed the same baseline system in terms of overall ranking.

Number of models used Contrary to the previous task edition, we find that more teams opted for the use of a single unified model for both datasets (27 out of 53), as opposed to using one model for each dataset. This is likely a result of a significant increase in the use of Large Language Models, an unsurprising shift that reflects the current research landscape in Natural Language Processing. Interestingly, the top four ranked teams can all be seen to adopt a 2-model approach, indicative of the potential benefits of having a distinct model specifically catering to the different lay summary styles of each dataset.

Use of additional data As with this previous task edition, we found that very few teams opted to make use of additional data (i.e., data not provided by the organizers as part of the task) in model development. As shown by the + column in Table 2,

[3]For these metrics, the scores are estimates of the US Grade level of education required to comprehend a given text.

[4]This represents a minor change from the averaging protocol of the previous task edition, in which we first calculated rankings for each criterion, before summing these rankings to compute an overall rank.

[5]Information about model selection is derived from both system papers submitted by participants and a "system description" field included in the submission form on CodaBench.

⋆	Team	#	+	Relevance				Readability				Factuality	
				R-1	R-2	R-L	BertS	FKGL	DCRS	CLI	LENS	AlignS	SummaC
1	**UIUC_BioNLP**	2	×	48.55	**15.69**	**45.50**	**86.77**	11.75	9.34	13.36	52.85	80.04	73.38
2	**Ctyun AI**	2	×	47.96	15.46	44.94	86.66	12.44	9.67	14.15	51.09	82.72	74.80
3	**Saama Technologies**	2	×	47.85	15.45	44.97	86.70	11.36	9.10	13.15	51.90	77.83	72.68
4	**WisPerMed**	2	×	47.12	15.18	44.28	86.53	11.07	8.86	12.87	51.03	78.18	72.16
5	**cylaun**	1	×	47.39	14.55	44.45	85.61	**10.46**	9.33	12.64	41.69	75.26	78.44
6	**BART Baseline**	2	×	46.96	13.95	43.58	86.23	12.04	10.15	13.49	48.10	77.88	70.26
7	**AUTH**	1	✓	48.23	14.57	44.77	85.76	12.44	10.04	13.50	66.11	74.18	66.40
8	**maverick**	1	×	42.77	12.97	39.42	85.01	15.04	10.65	16.61	52.30	91.22	83.85
9	**Empress**	1	×	43.96	12.29	41.36	84.89	10.66	9.06	12.89	59.73	73.47	68.02
10	**eulerian**	1	×	40.35	11.66	37.10	84.51	14.80	10.76	16.53	48.46	91.73	85.38
11	**BioLay_AK_SS**	2	×	43.98	12.15	40.39	84.71	14.20	11.12	15.12	49.57	85.03	78.60
12	**HULAT-UC3M**	2	×	**48.72**	14.65	45.20	86.22	12.71	10.43	14.08	49.34	66.69	67.03
13	**Atif_Tanish**	1	×	43.82	11.96	41.01	84.84	10.61	9.12	12.86	60.14	72.92	67.12
14	**qwerty**	1	×	37.26	10.45	34.48	83.54	13.36	9.18	14.60	42.16	89.89	83.23
15	**Deakin**	2	×	48.22	14.20	44.41	85.83	14.46	10.76	15.48	63.91	74.57	61.80
16	**MDSCL**	2	×	42.56	13.01	39.35	85.20	14.01	10.78	15.92	63.05	81.50	71.54
17	**MDS-CL**	2	×	42.13	12.90	38.93	85.14	14.13	10.82	15.96	61.71	81.98	73.14
18	**elirf**	2	✓	48.15	13.66	43.09	85.95	13.61	10.86	14.66	48.02	78.21	60.66
19	**RAG-RLRC-LaySum**	2	×	46.24	13.04	42.37	85.29	12.68	10.43	14.41	59.26	71.28	66.29
20	**naive_bhais**	2	×	43.42	12.60	39.91	85.72	12.89	10.94	14.32	37.86	81.34	67.81
21	**MDS-CL**	1	×	42.31	11.05	39.22	85.62	11.93	9.23	13.25	74.67	71.52	56.55
22	**MDS-CL**	1	×	43.43	11.98	40.13	85.55	12.39	9.76	14.28	76.80	72.18	54.41
23	**DhruvShlo**	1	×	42.15	11.05	39.40	84.42	11.76	9.08	13.02	49.17	71.25	63.98
24	**naman_tejas**	1	×	39.54	11.06	36.73	84.25	12.29	9.20	13.58	50.44	75.68	68.10
25	**SINAI**	2	×	42.05	12.49	38.53	85.83	12.23	9.86	13.81	76.95	71.17	53.98
26	**XYZ**	2	×	41.04	9.93	38.01	85.50	11.02	9.37	13.00	**81.21**	70.18	54.63
27	**gpsigh**	2	×	33.60	9.18	30.97	82.97	15.69	9.30	15.17	42.06	91.28	82.11
28	**YXZ**	2	×	42.25	10.91	39.20	84.99	11.18	8.57	12.44	71.57	64.89	53.49
29	**sanika**	2	×	42.90	11.16	38.06	83.33	17.93	12.40	17.37	11.37	85.16	**90.28**
30	**Bossy Beaver**	1	×	41.32	11.45	37.98	84.73	13.99	10.41	15.74	65.49	78.08	60.65
31	**Dayal K-Laksh G**	1	×	33.93	9.49	30.55	84.98	14.39	12.15	16.24	32.33	93.07	80.71
32	**MKGS**	1	×	37.75	9.72	34.67	83.33	15.79	11.92	17.50	22.07	**93.08**	83.52
33	**Shallow-Learning**	1	×	42.22	11.33	39.54	83.89	10.56	9.04	12.42	53.68	57.28	61.17
34	**NLPSucks**	2	×	34.91	8.32	33.32	82.62	10.68	**6.76**	12.08	37.86	74.36	64.22
35	**CookieMonster**	2	×	43.06	10.33	39.83	84.57	12.04	9.37	13.18	49.53	63.63	59.19
36	**NoblesseUranium**	1	×	39.16	10.34	35.87	84.65	14.21	10.44	15.45	51.99	75.74	67.32
37	**roon**	2	×	44.16	11.24	41.44	84.74	11.78	8.86	12.38	71.26	52.73	50.50
38	**jimmyapples**	2	×	43.36	10.84	40.35	84.82	11.44	9.03	12.10	71.48	56.54	49.00
39	**Shivam**	2	×	33.85	8.91	30.76	83.56	12.90	11.89	15.14	15.66	91.64	80.94
40	**HGP_NLP**	2	×	29.69	8.60	26.95	83.74	11.20	9.91	12.79	44.22	79.45	74.11
41	**Cornell-BioLay**	1	×	39.50	7.92	35.99	84.50	10.97	9.56	12.66	72.53	60.10	51.76
42	**xpc**	2	×	44.59	11.80	40.36	84.84	13.45	10.33	15.72	67.72	56.87	48.78
43	**Hemlo**	1	×	30.04	6.88	27.86	81.10	16.49	7.58	15.74	21.91	89.50	74.41
44	**anjaneya**	2	×	28.87	8.26	26.00	83.66	13.70	11.45	15.46	37.03	74.71	78.66
45	**Runtime_Terror**	1	×	40.18	10.14	37.44	83.69	13.98	8.41	13.13	49.60	47.51	50.37
46	**Abhi_Sidd**	1	×	35.33	9.15	31.79	83.09	17.29	12.42	14.47	17.41	79.09	62.95
47	**cbdch**	1	×	36.69	9.29	32.98	85.09	14.43	11.18	15.82	74.13	63.34	44.61
48	**aLoneLM**	1	×	35.67	6.74	33.29	82.33	11.07	8.52	**11.04**	42.82	48.93	52.62
49	**hohoho**	2	×	33.46	5.48	31.32	81.93	11.21	8.80	12.04	53.01	44.68	50.98
50	**huizige**	1	×	37.16	8.82	33.59	83.06	15.40	11.39	16.78	47.53	60.13	49.36
51	**SSS**	1	×	25.64	6.21	23.18	82.81	13.71	12.45	16.61	43.82	72.72	61.55
52	**H2P**	1	×	25.93	4.03	23.73	81.70	16.53	11.94	18.98	56.77	56.03	47.04
53	**KnowLab**	1	✓	32.16	7.34	28.31	80.63	36.29	11.55	11.28	1.32	42.41	54.74

Table 2: Task leaderboard - all metrics. The ⋆ column denotes the submission rank, the **#** column the number of models used - 1 (unified) or 2 (one for each dataset), and the **+** column the use of additional training data. **R** = ROUGE F1, **BertS** = BertScore, **FKGL** = Flesch-Kincaid Grade Level, **DCRS** = Dale-Chall Readability Score, **CLI** = Coleman-Liau Index, **AlignS** = AlignScore.

Rank	Team	Relevance				Readability				Factuality	
		R-1	R-2	R-L	BertS	FKGL	DCRS	CLI	LENS	AlignS	SummaC
1	**UIUC_BioNLP**	0.993	**1.000**	**1.000**	**1.000**	0.950	0.547	0.708	0.645	0.743	0.630
2	**Ctyun AI**	0.967	0.980	0.975	0.982	0.923	0.488	0.609	0.623	0.796	0.661
3	**Saama Technologies**	0.962	0.979	0.976	0.989	0.965	0.589	0.734	0.633	0.699	0.615
4	**WisPerMed**	0.931	0.956	0.945	0.962	0.976	0.631	0.770	0.622	0.706	0.603
5	**cylaun**	0.943	0.902	0.953	0.811	**1.000**	0.548	0.800	0.505	0.648	0.741
6	**BART Baseline**	0.924	0.851	0.914	0.913	0.939	0.405	0.693	0.586	0.700	0.561
7	**AUTH**	0.979	0.904	0.967	0.836	0.923	0.424	0.691	0.811	0.627	0.477
8	**maverick**	0.742	0.766	0.728	0.713	0.822	0.316	0.298	0.638	0.963	0.859
9	**Empress**	0.794	0.708	0.815	0.695	0.992	0.596	0.768	0.731	0.613	0.513
10	**eulerian**	0.637	0.655	0.624	0.632	0.832	0.298	0.308	0.590	0.973	0.893
11	**BioLay_AK_SS**	0.795	0.697	0.771	0.664	0.855	0.233	0.486	0.604	0.841	0.744
12	**HULAT-UC3M**	**1.000**	0.911	0.987	0.912	0.913	0.355	0.618	0.601	0.479	0.491
13	**Atif_Tanish**	0.788	0.680	0.799	0.686	0.994	0.585	0.771	0.736	0.602	0.493
14	**qwerty**	0.503	0.551	0.506	0.475	0.888	0.574	0.552	0.511	0.937	0.845
15	**Deakin**	0.978	0.872	0.951	0.848	0.845	0.298	0.441	0.784	0.635	0.376
16	**MDSCL**	0.733	0.770	0.725	0.745	0.862	0.294	0.385	0.773	0.771	0.590
17	**MDS-CL**	0.714	0.761	0.706	0.734	0.858	0.286	0.381	0.756	0.781	0.625
18	**elirf**	0.976	0.826	0.892	0.867	0.878	0.280	0.544	0.585	0.707	0.351
19	**RAG-RLRC-LaySum**	0.892	0.772	0.860	0.759	0.914	0.355	0.576	0.725	0.570	0.475
20	**naive_bhais**	0.790	0.749	0.773	0.850	0.912	0.270	0.578	0.498	0.760	0.493
21	**MDS-CL**	0.722	0.601	0.719	0.813	0.943	0.567	0.722	0.918	0.575	0.261
22	**MDS-CL**	0.771	0.682	0.759	0.802	0.925	0.473	0.592	0.945	0.588	0.214
23	**DhruvShlo**	0.716	0.602	0.727	0.618	0.950	0.593	0.751	0.599	0.569	0.424
24	**naman_tejas**	0.602	0.603	0.607	0.589	0.929	0.571	0.681	0.615	0.657	0.514
25	**SINAI**	0.711	0.726	0.688	0.848	0.932	0.455	0.651	0.947	0.568	0.205
26	**XYZ**	0.667	0.506	0.665	0.793	0.978	0.542	0.754	**1.000**	0.548	0.219
27	**gpsigh**	0.345	0.442	0.349	0.381	0.798	0.553	0.481	0.510	0.965	0.821
28	**YXZ**	0.720	0.590	0.718	0.711	0.972	0.682	0.825	0.879	0.444	0.194
29	**sanika**	0.748	0.612	0.667	0.440	0.711	0.008	0.203	0.126	0.844	**1.000**
30	**Bossy Beaver**	0.679	0.636	0.663	0.668	0.863	0.359	0.409	0.803	0.704	0.351
31	**Dayal K-Laksh G**	0.359	0.468	0.330	0.708	0.848	0.053	0.346	0.388	**1.000**	0.790
32	**MKGS**	0.525	0.488	0.515	0.440	0.794	0.093	0.187	0.260	**1.000**	0.852
33	**Shallow-Learning**	0.718	0.626	0.733	0.531	0.996	0.600	0.826	0.655	0.293	0.362
34	**NLPSucks**	0.402	0.368	0.454	0.324	0.991	**1.000**	0.870	0.457	0.631	0.429
35	**CookieMonster**	0.755	0.540	0.746	0.643	0.939	0.541	0.731	0.604	0.419	0.319
36	**NoblesseUranium**	0.586	0.541	0.568	0.656	0.855	0.353	0.445	0.634	0.658	0.497
37	**roon**	0.802	0.618	0.818	0.670	0.949	0.632	0.832	0.875	0.204	0.129
38	**jimmyapples**	0.768	0.584	0.769	0.683	0.962	0.601	0.867	0.878	0.279	0.096
39	**Shivam**	0.356	0.418	0.340	0.477	0.905	0.098	0.484	0.179	0.972	0.795
40	**HGP_NLP**	0.175	0.392	0.169	0.507	0.971	0.446	0.781	0.537	0.731	0.646
41	**Cornell-BioLay**	0.601	0.333	0.574	0.630	0.980	0.508	0.797	0.891	0.349	0.156
42	**xpc**	0.821	0.666	0.770	0.686	0.884	0.372	0.411	0.831	0.285	0.091
43	**Hemlo**	0.190	0.245	0.210	0.076	0.766	0.857	0.408	0.258	0.929	0.652
44	**anjaneya**	0.140	0.363	0.126	0.493	0.874	0.176	0.444	0.447	0.637	0.745
45	**Runtime_Terror**	0.630	0.524	0.639	0.499	0.864	0.711	0.737	0.604	0.101	0.126
46	**Abhi_Sidd**	0.420	0.439	0.386	0.400	0.735	0.005	0.568	0.201	0.724	0.401
47	**cbdch**	0.479	0.451	0.439	0.726	0.846	0.224	0.399	0.911	0.413	0.000
48	**aLoneLM**	0.435	0.233	0.453	0.277	0.976	0.691	**1.000**	0.520	0.129	0.175
49	**hohoho**	0.339	0.124	0.365	0.212	0.971	0.642	0.874	0.647	0.045	0.139
50	**huizige**	0.499	0.410	0.466	0.395	0.809	0.187	0.277	0.578	0.350	0.104
51	**SSS**	0.000	0.187	0.000	0.356	0.874	0.000	0.299	0.532	0.598	0.371
52	**H2P**	0.013	0.000	0.025	0.174	0.765	0.089	0.000	0.694	0.269	0.053
53	**KnowLab**	0.283	0.283	0.230	0.000	0.000	0.158	0.971	0.000	0.000	0.222

Table 3: Task leaderboard with min-max normalization. **R** = ROUGE F1, **BertS** = BertScore, **FKGL** = Flesch-Kincaid Grade Level, **DCRS** = Dale-Chall Readability Score, **CLI** = Coleman-Liau Index, **AlignS** = AlignScore.

only three teams - **AUTH**, **elirf**, and **KnowLab** - indicated that they adopted such an approach.

Reflection on evaluation protocol changes Here, we discuss the impact of the changes made to the evaluation protocol over the previous task edition. As mentioned in §4, the first of these changes surrounds the introduction of new metrics for the Readability and Factuality criteria. As a model-based simplification metric, LENS was introduced to provide an additional angle for teams to consider for Readability, with Maddela et al. (2023) demonstrating that the metric correlates particularly well with the *fluency* ratings of human annotators for simplified texts. Notably, LENS does not exhibit a strong alignment with other (more heuristic) Readability metrics, suggesting that these metrics may not capture this aspect of simplified texts.

For Factuality, we introduced the AlignScore and SummaC metrics as a replacement for a fine-tuned version of BARTScore to avoid potential bias toward BART-based models. However, given that these metrics broadly involve comparing a generated summary to the source text, these metrics tend to favor highly extractive outputs. Given that reference lay summaries tend to be quite abstractive (particularly in the case of the eLife dataset), this resulted in a trade-off between scoring highly for Factuality and the metrics of Relevance or Readability. Overall, we observe that the systems that ranked the highest were those that most successfully balanced this trade-off, typically obtaining strong Relevance and Readability scores while maintaining relatively high Factuality scores.

Finally, the process for the calculation of final rankings was changed from summing individual criterion rankings to the averaging of average criterion scores. This change was motivated by the failure of the previous method of ranking to take into account the relative difference between average scores for a given criterion, something that was commented on by last year's participants.[6] However, the new ranking system was also found to be not without its issues, particularly surrounding the existence of outliers. Specifically, it was observed that, if there existed teams that scored particularly poorly for a given metric, then all other teams would obtain relatively strong (and less diverse) scores for this metric relative to others - this can be seen for the FKGL metric in Table 3.

[6]For example, in terms of average criterion score, the team ranked 1st may outperform the team ranked 2nd by a large margin, who in term may outperform the team ranked 3rd by a small margin. However, by converting these scores to rankings, all differences are treated as equal.

6 Submissions

Out of the 53 participating teams, 14 teams submitted system papers. Here, we provide a brief summary of the approaches taken by these teams.

UIUC_BioNLP (You et al., 2024) This team produced the top-ranked submission, adopting an extract-then-summarize approach that utilizes TextRank (Mihalcea and Tarau, 2004) for salient sentence extraction, followed by a fine-tuned GPT-3.5-turbo model for summary generation. Specifically, their submitted system extracted the top 40 most salient sentences using TextRank, and their GPT-bsaed model is fine-tuned on 200 examples. Additional experimentation was conducted using various extractive summarization approaches and comparing the number of examples required for effective fine-tuning. Furthermore, the team also explored a LongFormer-based approach that further incorporates retrieved Wikipedia data in a Retrieval-Augmented Generation (RAG) setup.

Cytun AI (Zhao et al., 2024) Making the second-ranked submission, the methodology of this team surrounds the use of fine-tuned LLMs. As part of their experimentation, they compare two approaches for handling lengthy input articles: hard truncation and text chunking. Additionally, their summary-generation pipeline includes data preprocessing, augmentation, and prompt engineering.

Saama Technologies (Kim et al., 2024) This team achieved the third-ranking submission, which surrounded fine-tuning a Mistral-7B model[7] in an unsupervised fashion using low-ranked adaptation (LoRA) (Hu et al., 2021), followed by zero-shot summary generation and post-processing to remove redundant sentences. This team also experiments with several other fine-tuning methods, including supervised fine-tuning with LoRA and Direct Preference Optimization (Rafailov et al., 2023).

WisPerMed (Pakull et al., 2024) Ranking in fourth place, the selected submission of WisPerMed utilized a fine-tuned BioMistral model, combined with few-shot prompting and a Dynamic-Expert selection (DES) mechanism. Specifically, their BioMistral Model was trained using abstracts and lay summaries of the provided train set; and

[7]mistral-7B-instruct-v0.2

their proposed DES mechanism involved generating several lay summary versions with different prompts for a given input, before selecting the most desirable based on the scores of the references-less Readability and Factuality metrics used in the task. In additional experiments, they also measured system performance utilizing LLAMA3, as well as that of few-shot and zero-shot model variants. The task organizers selected this team to receive an award for the "most innovative approach".

AUTH (Stefanou et al., 2024) Being one of the only teams to utilize external data, this retrieves 300 abstract-lay abstract pairs scraped from the Science Journal for Kids website.[8] They use this retrieved data as in-context examples for GPT-4, which they prompt to augment the provided datasets by rewriting reference summaries with higher readability scores. Finally, they use this data to fine-tune to fine-tune BioBART (Yuan et al., 2022), whilst also experimenting with controllable generation techniques in the form of control tokens prepended to the input article (`<elife>` / `<plos>` and `<general_lay_summary>` / `<kids_lay_summary>`).

Eulerian (Modi and Karthikeyan, 2024) The team experimented with different combinations of the FLAN-T5 (Chung et al., 2022b) model variations and data selection. They compare the performance of these methods with a preposed "Preprocessing over Abstract" technique, in which they use a regular expression to remove some abstract information (i.e., anything inside of parentheses, braces and brackets), finding that this outperforms all neural methods tested in terms of Relevance and Factuality metrics.

BioLay_AK_SS (Karotia and Susan, 2024) Focusing largely on data augmentation, this team generated additional summary samples using 2 general-purpose models: BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020a). The augmented dataset was then used to fine-tune a domain-specific BioBART model, which was found to improve its overall improved overall performance.

HULAT-UC3M (Gonzalez and Martínez, 2024) Again comparing the performance of domains-specific and general-purpose models, this team experimented with fine-tuning both Longformer (Beltagy et al., 2020) and BioBART models on the given datasets. Additionally, they experiment with extending BioBART to utilise Longformer-based sparse attention, thus allowing it to process longer inputs. Overall, they found that fine-tuning the standard BioBART model on each dataset yields the best performance.

DeakinNLP (Quoc To et al., 2024) This team assessed the performance of both a fine-tuned Longformer and GPT-4 (with zero- and few-shot prompting). Additional analysis is also conducted surrounding data selection and the performance vs. cost trade-off between select methods.

elirf (Ahuir et al., 2024) Again utilising Longformer as their base model, this team experimented with domain-adaption via a continuous pre-training approach. During pre-training, several pretraining tasks were aggregated to inject linguistic knowledge and increase the abstractiveness of generated summaries. Finally, they developed a regression-based ranking model that improved system performance by selecting the most promising from a set of generated summaries.

RAG-RLRC-LaySum (Ji et al., 2024) This team developed a Retrieval-Augmented Generation (RAG) Lay Summarisation approach, utilizing multiple knowledge sources (including both source documents and Wikipedia). They experiment with LLMs (Gemini and ChatGPT) for both summary generation and paraphrasing, in addition to a Longformer baseline. Lastly, the team also develop a Reinforcement Learning strategy to fine-tune the readability of generated summaries.

SINAI (Chizhikova et al., 2024) Focusing largely on a few-shot setting, this team compared the performance of several popular LLMs including GPT-3.5, GPT-4, and LLAMA3. Further experimentation surrounded the fine-tuning of a smaller LLAMA model (LLAMA3-8B) using both parameter-efficient LoRA techniques and Direct Preference Optimization (Rafailov et al., 2023).

XYZ (Zhou et al., 2024) This team performed a thorough comparison of several state-of-the-art LLMs, focusing largely on comparing the readability of generated summaries. Further experimentation surrounds Summary rewriting, Title infusing, K-shot prompting, and LoRA-based fine-tuning, with their best-performing submission utilizing a combination of these methods and obtaining the best overall Readability scores.

[8] https://sciencejournalforkids.org/

HGP_NLP (Malik et al., 2024) This team fine-tune and evaluate multiple T5 model variants, also experimenting with LoRA-based fine-tuning.

7 Conclusion

The second edition of the BioLaySumm Shared Task was hosted by the BioNLP Workshop@ACL 2024. Several changes were implemented over the previous edition of the task covering participation rules, evaluation metrics, and ranking protocol. In terms of participant engagement, the task attracted a total of 53 teams, representing a significant growth from the previous edition's 20 teams. Our results indicate a drastic shift towards the use of LLMs for lay summarisation, with a wide range of both domain-specific and general-purpose LLMs being adopted in various settings across participant submissions.

References

Vicent Ahuir, Diego Torres, Encarna Segarra, and Lluís-F Encarna. 2024. Elirf-vrain at biolaysumm: Boosting lay summarization systems performance with ranking models. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

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

Mariia Chizhikova, Manuel Carlos Díaz-Galiano, L. Alfonso Ureña-López, and María-Teresa Martín-Valdivia. 2024. Sinai at biolaysumm: Self-play fine-tuning of large language models for biomedical lay summarisation. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022a. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022b. Scaling instruction-finetuned language models.

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.

Adrian Gonzalez and Paloma Martínez. 2024. Hulat-uc3m at biolaysumm: Adaptation of biobart and longformer models to summarizing biomedical documents. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

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

Yuelyu Ji, Zhuochun Li, Rui Meng, Sonish Sivarajkumar, Yanshan Wang, Zeshui Yu, Hui Ji, Yushui Han, Hanyu Zeng, and Daqing He. 2024. Rag-rlrc-laysum at biolaysumm: Integrating retrieval-augmented generation and readability control for layman summarization of biomedical texts. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

Akanksha Karotia and Seba Susan. 2024. Biolay_ak_ss at biolaysumm: Domain adaptation by two-stage fine-tuning of large language models used for biomedical lay summary generation. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

Hwanmun Kim, Kamal raj Kanakarajan, and Malaikannan Sankarasubbu. 2024. Saama technologies at biolaysumm: Abstract based fine-tuned models with lora. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains.

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.

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*.

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.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Hemang Malik, Gaurav Pradeep, and Pratinav Seth. 2024. Hgp-nlp at biolaysumm: Leveraging lora for lay summarization of biomedical research articles using seq2seq transformers. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

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.

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

Tabea Margareta Grace Pakull, Hendrik Damm, Ahmad Idrissi-Yaghir, Henning Schäfer, Peter A. Horn, and Christoph M. Friedrich. 2024. Wispermed at biolaysumm: Adapting autoregressive large language models for lay summarization of scientific articles. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

Huy Quoc To, Ming Liu, and Guangyan Huang. 2024. Deakinnlp at biolaysumm: Evaluating fine-tuning longformer and gpt-4 prompting for biomedical lay summarization. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Loukritia Stefanou, Tatiana Passali, and Grigorios Tsoumakas. 2024. Auth at biolaysumm 2024: Bringing scientific content to kids. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

Zhiwen You, Shruthan Radhakrishna, Shufan Ming, and Halil Kilicoglu. 2024. Uiuc_bionlp at biolaysumm: an extract-then-summarize approach augmented with wikipedia knowledge for biomedical lay summarization. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

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

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

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.

Ruijing Zhao, Siyu Bao, Siqin Zhang, Jinghui Zhang, Weiyin Wang, and Yunian Ru. 2024. Ctyun ai at

biolaysumm: Enhancing lay summaries of biomedical articles through large language models and data augmentation. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

Jieli Zhou, Cheng Ye, Pengcheng Xu, and Hongyi Xin. 2024. Team yxz at biolaysumm: Adapting large language models for biomedical lay summarization. In *Proceedings of the 23rd Workshop on Biomedical Language Processing*, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

⋆	Team	Relevance	Readability	Factuality	Avg.
1	**UIUC_BioNLP**	0.998	0.712	0.686	0.799
2	**Ctyun AI**	0.976	0.661	0.728	0.788
3	**Saama Technologies**	0.977	0.730	0.657	0.788
4	**WisPerMed**	0.948	0.750	0.655	0.784
5	**cylaun**	0.902	0.713	0.694	0.770
6	**BART Baseline**	0.901	0.655	0.631	0.729
7	**AUTH**	0.922	0.713	0.552	0.729
8	**maverick**	0.737	0.519	0.911	0.723
9	**Empress**	0.753	0.772	0.563	0.696
10	**eulerian**	0.637	0.507	0.933	0.692
11	**BioLay_AK_SS**	0.732	0.545	0.793	0.690
12	**HULAT-UC3M**	0.952	0.622	0.485	0.686
13	**Atif_Tanish**	0.738	0.772	0.547	0.686
14	**qwerty**	0.509	0.631	0.891	0.677
15	**Deakin**	0.912	0.592	0.506	0.670
16	**MDSCL**	0.743	0.579	0.681	0.667
17	**MDS-CL**	0.729	0.570	0.703	0.667
18	**elirf**	0.890	0.572	0.529	0.664
19	**RAG-RLRC-LaySum**	0.821	0.643	0.522	0.662
20	**naive_bhais**	0.790	0.565	0.626	0.660
21	**MDS-CL**	0.714	0.788	0.418	0.640
22	**MDS-CL**	0.753	0.734	0.401	0.629
23	**DhruvShlo**	0.666	0.723	0.497	0.628
24	**naman_tejas**	0.600	0.699	0.585	0.628
25	**SINAI**	0.743	0.746	0.386	0.625
26	**XYZ**	0.658	0.819	0.384	0.620
27	**gpsigh**	0.379	0.585	0.893	0.619
28	**YXZ**	0.685	0.840	0.319	0.614
29	**sanika**	0.617	0.262	0.922	0.600
30	**Bossy Beaver**	0.662	0.609	0.528	0.599
31	**Dayal K-Laksh G**	0.466	0.409	0.895	0.590
32	**MKGS**	0.492	0.333	0.926	0.584
33	**Shallow-Learning**	0.652	0.769	0.328	0.583
34	**NLPSucks**	0.387	0.830	0.530	0.582
35	**CookieMonster**	0.671	0.704	0.369	0.581
36	**NoblesseUranium**	0.588	0.572	0.577	0.579
37	**roon**	0.727	0.822	0.166	0.572
38	**jimmyapples**	0.701	0.827	0.187	0.572
39	**Shivam**	0.398	0.417	0.884	0.566
40	**HGP_NLP**	0.311	0.684	0.688	0.561
41	**Cornell-BioLay**	0.535	0.794	0.253	0.527
42	**xpc**	0.736	0.625	0.188	0.516
43	**Hemlo**	0.180	0.572	0.791	0.514
44	**anjaneya**	0.281	0.485	0.691	0.486
45	**Runtime_Terror**	0.573	0.729	0.113	0.472
46	**Abhi_Sidd**	0.411	0.378	0.563	0.450
47	**cbdch**	0.524	0.595	0.207	0.442
48	**aLoneLM**	0.349	0.797	0.152	0.433
49	**hohoho**	0.260	0.783	0.092	0.378
50	**huizige**	0.443	0.463	0.227	0.378
51	**SSS**	0.136	0.426	0.484	0.349
52	**H2P**	0.053	0.387	0.161	0.200
53	**KnowLab**	0.199	0.282	0.111	0.197

Table 4: Task leaderboard with min-max normalization. The ⋆ column denotes the submission rank. **R** = ROUGE F1, **BertS** = BertScore, **FKGL** = Flesch-Kincaid Grade Level, **DCRS** = Dale-Chall Readability Score, **CLI** = Coleman-Liau Index, **AlignS** = AlignScore.