# Overview of the BioLaySumm 2025 Shared Task on Lay Summarization of Biomedical Research Articles and Radiology Reports

**Chenghao Xiao**[1], **Kun Zhao**[2], **Xiao Wang**[3], **Siwei Wu**[3], **Sixing Yan**[4], **Tomas Goldsack**[5], **Sophia Ananiadou**[3], **Noura Al Moubayed**[1], **Zhan Liang**[2], **William Cheung**[4], **Chenghua Lin**[3]
[1]Durham University, [2]University of Pittsburgh, [3]The University of Manchester
[4]Hong Kong Baptist University, [5] The University of Sheffield
chenghao.xiao@durham.ac.uk chenghua.lin@manchester.ac.uk

## Abstract

This paper presents the setup and results of the third edition of the BioLaySumm shared task on Lay Summarization of Biomedical Research Articles and Radiology Reports, hosted at the BioNLP Workshop at ACL 2025. In this task edition, we aim to build on the first two editions' successes 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. Specifically, we introduce the new task of Radiology Report Generation with Layman's terms, which is parallel to the task of lay summarization of biomedical articles in the first two editions. Overall, our results show that a broad range of innovative approaches were adopted by task participants, including inspiring explorations of latest RL techniques adopted in the training of general-domain large reasoning models.

## 1 Introduction

Lay Summarization describes the task of transforming a technical or specialist text that into summaries accessible to non-expert audience. By prioritizing clarity, context, and relevance over specialized terminologies, lay summaries bridge critical knowledge gaps between experts and diverse stakeholders, including practitioners, researchers in adjacent fields, patients, and the public. Despite their value in democratizing information, the creation of high-quality lay summaries remains scarce and labour-intensive, creating significant barriers to inclusive knowledge dissemination.

The need for accessible communication spans the entire biomedical ecosystem, from cutting-edge research to routine clinical care. Biomedical research publications, which contain the latest findings on prominent health-related topics, represent a key area where lay summarization is crucial. While mandatory for some journals, lay summaries are not universally adopted, leaving vital research inaccessible to non-experts. Even when required, authors who are often untrained in science communication struggle to distill their work effectively. Automatic lay summarization thus offers immense potential to scale accessibility while alleviating authorial burden, ensuring findings reach patients, policymakers, and interdisciplinary researchers

Parallel challenges exist in **clinical communication**, particularly in **radiology**. The 21st Century Cures Act (21st Century Cures Act, 2016) mandates immediate patient access to electronic health records, yet radiology reports—designed for clinicians—use highly technical language. Fewer than 4% radiology reports meet the eighth-grade reading level typical of U.S. adults (Martin-Carreras et al., 2019), causing confusion, anxiety, and poor adherence to follow-up care. Creating lay summaries of these reports is therefore not just a matter of convenience but a critical step toward a more patient-centered, transparent, and effective healthcare system.

The BioLaySumm shared task[1] is dedicated to advancing the automatic lay summarization of biomedical texts. Building on the success of the first two editions (Goldsack et al., 2023, 2024), **this year's shared task** addresses two domains: **biomedical articles** and **radiology reports**. Through this shared task, we aim to encourage the development of novel approaches and increase research interest in developing techniques for making scientific and clinical information accessible to broader audiences. In this paper, we present the results of the third edition of the BioLaySumm shared task, hosted by the BioNLP Workshop at ACL 2025. This year, we expand the scope of our challenge to include two parallel tracks: (i) the established task of Lay Summarization of Biomedical Research Articles; and (ii) a new track on the Lay Summarization of Radiology Reports.

In what remains of the paper, we address the formulation of these two tasks (§2), the datasets

[1]https://biolaysumm.org



*Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025)*, pages 365–377

used (§3), and the evaluation procedure (§4), before providing a description of the participating systems (§5), and notable insights (§6).

## 2 Task Description

As part of the BioLaySumm 2025 shared task, participants developed systems capable of generating accessible summaries of biomedical content for non-expert audiences. Building upon previous editions, this year's competition introduced new challenges while maintaining core evaluation frameworks. The task was hosted using the CodaBench platform (Xu et al., 2022), with submissions automatically evaluated upon upload.

### 2.1 Task 1: Lay Summarization of Biomedical Articles

In Task 1, participants were required to generate plain-language summaries from technical research articles, with two distinct subtasks:

**Subtask 1.1: Plain Lay Summarization** required generating summaries using only the article's abstract and main text as input. As in previous editions, two separate datasets (**PLOS** and **eLife**) with notable stylistic differences were provided. Systems could employ either:

- Separate models trained independently on each dataset
- A unified model trained on both datasets

Final rankings were determined by average performance across both datasets.

**Subtask 1.2: Lay Summarization with External Knowledge** extended the plain summarization task by mandating incorporation of external resources to address knowledge gaps for lay audiences. Participants employed techniques such as Retrieval-Augmented Generation (RAG) or manual augmentation to integrate supplementary information (e.g., background context, terminology definitions).

### 2.2 Task 2: Radiology Report Generation

New in 2025, this task focused on translating medical imaging reports into patient-friendly explanations:

**Subtask 2.1: Radiology Report Translation** involved text-to-text simplification of professional radiology reports. Participants utilized report-layman term pairs from multiple datasets (Open-i, PadChest, BIMCV-COVID19 ± MIMIC-CXR), with separate rankings for systems using three versus four datasets.

**Subtask 2.2: Multimodal Translation** (optional) required generating lay summaries directly from medical images using end-to-end models (e.g., multimodal LLMs), with separate evaluation tracks based on training data scope.

**Competition Framework** Consistent with previous editions:

- Participants received training/validation sets with reference summaries alongside blind test sets
- For text-only tasks (Task 1 and Subtask 2.1), llama3 8B/Qwen2.5 7B will be used as the primary baseline.
- For multimodal task (Subtask 2.2), we will use finetuned Qwen-VL 7B as the finetuned baseline.

Detailed dataset characteristics appear in §3, with evaluation protocols in §4. Participants could attempt any combination of subtasks based on their research interests.

## 3 Datasets

The datasets used for the Task 1 are based on the previous works of Goldsack et al. (2022) and Luo et al. (2022), 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

| Dataset | Task | # Train | # Val | # Test |
|---|---|---|---|---|
| eLife | 1 | 4,346 | 241 | 142 |
| PLOS | 1 | 24,773 | 1,376 | 142* |
| PadChest | 2 | 116,847 | 7,824 | 7,130 |
| BIMCV-COVID19 | 2 | 31,364 | 2,042 | 3,221 |
| Open-i | 2 | 2,243 | 1,34 | 186 |
| MIMIC-CXR | 2 | 45,000 | 5,000 | 500 |

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

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).

For Task 2, we utilized four radiology datasets: PadChest (Bustos et al., 2020), BIMCV-COVID19+ (Vayá et al., 2020), Open-i (Demner-Fushman et al., 2012), and MIMIC-CXR (Johnson et al., 2019). The PadChest dataset comprises over 160,000 images from 67,000 patients, interpreted by radiologists at San Juan Hospital (Spain) between 2009 and 2017, and includes six positional views with supplementary acquisition and demographic metadata. The BIMCV-COVID19+ dataset contains chest X-rays (CXR/DX) and computed tomography (CT) images of COVID-19 patients, accompanied by radiographic findings, pathologies, polymerase chain reaction (PCR) tests, immunoglobulin G (IgG)/M (IgM) antibody tests, and reports from the Valencian Community Medical Image Database (BIMCV). This database includes 21,342 CR, 34,829 DX, and 7,918 CT studies. Open-i offers access to 3.7 million images from 1.2 million PubMed Central articles, including 7,470 chest X-rays with 3,955 reports. The MIMIC-CXR dataset contains 377,110 JPEG images with structured labels derived from 227,827 associated free-text reports, de-identified to comply with HIPAA Safe Harbor requirements by removing protected health information (PHI).

For the layman-style reports of Task 2, we applied the method from Zhao et al. (2025) to create the layman-style reports for all four datasets. PadChest and BIMCV-COVID19+ reports were first translated into English before transformation; Open-i and MIMIC-CXR were converted directly. A subset of MIMIC-CXR reports was selected for training and testing in this shared task.

Table 1 summarizes the data split information for all datasets of two Tasks. Note that the training and validation sets used for both datasets are identical to those published in Goldsack et al. (2022) and Zhao et al. (2025). By leveraging these datasets, we aim to develop abstractive summarization models and layman-style report generation systems capable of producing accessible summaries for unseen biomedical articles and layman-style radiology reports. This approach will facilitate effective communication of significant new publications to non-expert audiences and patients across diverse biomedical domains.

## 4 Evaluation

**Task1: Lay Summarization** For both subtasks of Task 1, we evaluate summary quality according to three criteria - *Relevance*, *Readability*, and *Factuality* - where each criterion is composed of one or more automatic metrics:

- *Relevance*: ROUGEROUGE - 1, 2, and L (Lin, 2004), *BLEU (Papineni et al., 2002), *METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020).
- *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)

**Task2: Radiology Report Generation** For both subtasks of Task 2, we evaluate report quality according to three criteria - *Relevance*, *Readability*, and *Clinical* - where each criterion is composed of one or more automatic metrics:

- *Relevance*: ROUGE - 1, 2, and L (Lin, 2004), BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020), *Semantic Similarity scores (Pesquita et al., 2009).
- *Readability*: Flesch-Kincaid Grade Level (FKGL), Dale-Chall Readability Score (DCRS), Coleman-Laiu Index (CLI).
- *Clinical Metrics*: *CheXbert (Smit et al., 2020), and *RadGraph (Jain et al., 2021).

Here "*" indicates that the metric is newly introduced for this year's edition of the task. Specifically, the BLEU and METEOR metrics are introduced to measure how closely a system-generated

summary or report matches its reference at the lexical level. To assess report quality at the semantic level, we introduce a semantic score measured based on the cosine similarity between the sentence-level embeddings of each generated report and its reference. Additionally, CheXbert and RadGraph are introduced to quantify clinical correctness, which can not be assessed by general metrics. By incorporating these two domain-aware metrics, the evaluation process could be more comprehensive.

For Task 1, The scores calculated for each metric are the average of those calculated independently for the generated lay summaries of PLOS and eLife. As for Task 2 (open track), all scores are computed on the combined public datasets — PadChest, BIMCV-COVID19, and Open-i. While for Task 2 (closed track), the reports are evaluated on open-track datasets plus MIMIC-CXR, and each metric is then averaged over the two scores.

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.[2]

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. 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.

**Baselines** We train Qwen2.5 (Qwen et al., 2025), LLaMA3 (Grattafiori et al., 2024), and Qwen2.5-VL (Bai et al., 2025) on the BioLaySumm 2025 training dataset as the baseline models. (1) For **Task 1 (Lay Summarization)**, we select Qwen2.5-7B-Instruct and LLaMA3-8B-Instruct as the backbone models and train them on our training data by using the whole article as input and the lay summary as output. (2) For the **Task 2 (Radiology Report Generation with Layman's Terms)**, we train Qwen2.5-7B-Instruct and LLaMA3-8B-Instruct on our training data for Subtask 2.1 (Radiology Report Translation) and train Qwen2.5-VL-7B-Instruct for the Subtask 2.2 (Multi-modal Radiology Report Translation).

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

## 5 Submissions

Out of all participating teams, 13 teams submitted system papers. Here, we provide a brief summary of the approaches taken by these teams.

**AEHRC** (Zhang et al., 2025) This team produced the top-ranked submission for both open-source and close-source tracks of Subtask 2.1, and provided a comparison study between encoder-decoder and decoder-only architectures. The paper presents the surprising results that a 700M T5-large-based model provides better performance than a 3B LLaMA-3.2-based model across nine out of ten metrics, including relevance, readability, and clinical accuracy, despite having significantly fewer parameters. The findings highlight the continual relevance of encoder-decoder models for lay summarization tasks in the era of LLMs.

**MetninOzu** (Evgin et al., 2025) This team proposes two innovative approaches, reverse data augmentation and salient sentence injection, and a detailed study of them. The authors curated a dataset of child-friendly articles with corresponding gold-standard summaries and used LLMs to rewrite them into more complex scientific variants to augment the training data beyond the shared-task training set. They also investigated whether they can insert salient sentences from the main article into the summary to enrich the input, leveraging sentence embedding models.

**XSZ** (Xu et al., 2025) This team investigates (i) k-shot demonstration fine-tuning with LLMs, and (ii) further employing latest reasoning-oriented RL methods to LLMs. For the first method, they use embedding models to retrieve top-K examples and fine-tune a Llama3-8B with LoRA. They then employ RL algorithms (PPO and GRPO) to further fine-tune the models. The reward function is specifically design to optimize the evaluated metrics, including factual metrics, relevant metrics and readability metrics. Although the RL results are not submitted to the competition, the paper is well-implemented and innovative, showing that RL methods are useful for lay summarization.

| Rank | Team | Relevance | | | | Readability | | | | Factuality | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ROUGE | BLEU | MTR | BERTS | FKGL | DCRS | CLI | LENS | AlignS | SummaC |
| 1 | **SUWMIT** | **0.370** | **10.07** | 0.308 | 0.864 | 11.74 | 9.08 | 12.58 | 72.61 | 0.750 | 0.682 |
| 2 | **Baseline-llama3-8B-sft** | 0.366 | 9.86 | **0.314** | 0.863 | 12.20 | 9.25 | 12.98 | 72.86 | 0.722 | 0.644 |
| 3 | **Baseline-qwen2.5-7B-sft** | 0.352 | 8.74 | 0.303 | 0.870 | 12.71 | 9.65 | 13.70 | 60.22 | 0.754 | 0.644 |
| 4 | **BDA-UCM** | 0.334 | 8.08 | 0.294 | **0.870** | 12.32 | 9.26 | 13.20 | 64.07 | 0.691 | 0.590 |
| 5 | **MetinOZU** | 0.330 | 6.95 | 0.290 | 0.857 | 16.45 | 11.22 | 17.01 | 34.86 | **0.881** | **0.920** |
| 6 | **MIRAGES** | 0.288 | 4.63 | 0.230 | 0.846 | 11.71 | **8.46** | 11.99 | 71.27 | 0.681 | 0.605 |
| 7 | **TupiQ** | 0.335 | 7.16 | 0.268 | 0.862 | 13.44 | 10.59 | 13.48 | 43.67 | 0.762 | 0.642 |
| 8 | **LaySummX** | 0.321 | 5.44 | 0.253 | 0.855 | 12.33 | 9.51 | 13.38 | 80.46 | 0.675 | 0.521 |
| 9 | **CUTN_Bio** | 0.268 | 3.25 | 0.226 | 0.848 | **10.52** | 8.84 | **11.43** | **84.14** | 0.589 | 0.549 |
| 10 | **Aard** | 0.319 | 5.45 | 0.293 | 0.851 | 14.56 | 10.02 | 15.36 | 71.51 | 0.695 | 0.509 |
| 11 | **LTRC** | 0.288 | 4.27 | 0.222 | 0.850 | 13.36 | 9.30 | 13.29 | 79.34 | 0.601 | 0.476 |
| 12 | **5cNLP** | 0.333 | 6.14 | 0.268 | 0.859 | 16.07 | 10.40 | 15.34 | 76.05 | 0.631 | 0.549 |
| 13 | **RainCityNLP** | 0.284 | 4.87 | 0.241 | 0.840 | 16.74 | 11.66 | 16.24 | 9.41 | 0.612 | 0.653 |
| 14 | **SXZ** | 0.165 | 1.33 | 0.153 | 0.801 | 12.59 | 11.83 | 13.29 | 6.56 | 0.862 | 0.528 |
| 15 | **demo** | 0.165 | 1.33 | 0.153 | 0.801 | 12.59 | 11.83 | 13.29 | 6.56 | 0.862 | 0.528 |
| 16 | **x2z** | 0.182 | 1.18 | 0.168 | 0.804 | 12.60 | 8.56 | 12.65 | 63.22 | 0.368 | 0.468 |

(a) SubTask 1.1: Plain Lay Summarization

| Rank | Team | Relevance | | | | Readability | | | | Factuality | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ROUGE | BLEU | MTR | BERTS | FKGL | DCRS | CLI | LENS | AlignS | SummaC |
| 1 | **Aard** | 0.292 | 4.32 | 0.262 | 0.848 | **11.16** | **8.36** | **11.94** | **81.50** | 0.614 | **0.537** |
| 2 | **CUTN_Bio** | 0.296 | 4.08 | 0.228 | 0.855 | 13.37 | 10.25 | 14.74 | 80.00 | **0.689** | 0.507 |
| 3 | **5cNLP** | **0.335** | **5.91** | **0.275** | **0.858** | 16.30 | 10.29 | 15.24 | 75.57 | 0.610 | 0.445 |
| 4 | **LTRC** | 0.215 | 2.01 | 0.169 | 0.818 | 13.71 | 9.66 | 13.60 | 74.48 | 0.378 | 0.429 |

(b) Subtask 1.2: Lay Summarization with External Knowledge

| Rank | Team | Relevance | | | | | Readability | | | Clinical Metrics | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ROUGE | BLEU | MTR | BERTS | SIM | FKGL | DCRS | CLI | CHEX | RG |
| 1 | **AEHRC** | **0.671** | **46.09** | **0.704** | **0.953** | **0.890** | **7.397** | 9.31 | **8.05** | **0.860** | **0.402** |
| 2 | **KHU_LDI** | 0.529 | 28.66 | 0.577 | 0.935 | 0.843 | 7.528 | **9.29** | 8.26 | 0.827 | 0.265 |

(c) Subtask 2.1: Radiology Report Translation (Open Track)

| Rank | Team | Relevance | | | | | Readability | | | Clinical Metrics | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ROUGE | BLEU | MTR | BERTS | SIM | FKGL | DCRS | CLI | CHEX | RG |
| 1 | **AEHRC** | **0.629** | **38.99** | **0.669** | **0.948** | **0.894** | 7.574 | 8.97 | 7.95 | 0.777 | **0.377** |
| 2 | **Baseline-qwen2.5-7B-sft** | 0.537 | 25.71 | 0.543 | 0.938 | 0.854 | **6.440** | 10.04 | 8.55 | 0.779 | 0.291 |
| 3 | **5cNLP** | 0.555 | 28.27 | 0.609 | 0.937 | 0.872 | 8.046 | 9.24 | 8.23 | 0.750 | 0.317 |
| 4 | **Baseline-llama3-8B-sft** | 0.527 | 25.18 | 0.527 | 0.936 | 0.847 | 6.785 | **8.53** | 8.67 | **0.806** | 0.286 |
| 5 | **CUTN_Bio** | 0.404 | 14.90 | 0.428 | 0.913 | 0.798 | 7.359 | 8.53 | **7.36** | 0.704 | 0.216 |

(d) Subtask 2.1: Radiology Report Translation (Closed Track)

Table 2: Task leaderboard - all metrics. **BertS** = BertScore, **FKGL** = Flesch-Kincaid Grade Level, **DCRS** = Dale-Chall Readability Score, **CLI** = Coleman-Liau Index, **MTR** = METOR, **SIM** = Similarity, **AlignS** = AlignScore, **CHEX** = F1 chexbert, **RG** = Radgraph.

**Aard** (Gupta and Krishnamurthy, 2025) This team introduced a modular and flexible system designed for generating lay summaries by leveraging large language models, a BioBERT-based named entity recognizer, and the UMLS knowledge base. For Task 1.1, they focused on summarization using only the internal content of articles, while Task 1.2 enhanced this with external biomedical knowledge like terminology definitions to improve readability and factuality. Their approach involved chunking articles, extracting key sentences, iterative rewriting, and integrating simplified definitions for complex terms. The LayForge system demonstrated strong performance, especially in readability metrics, highlighting the effectiveness of domain-specific augmentation for lay summary generation.

**RainCityNLP** (Wilson et al., 2025) This team utilized TF-IDF for sentence scoring and experimented with Pegasus-XSum and a Falcons.ai model pre-trained on medical data. All experiments were conducted on consumer-grade hardware, demonstrating feasibility in low-resource settings. Evaluation showed the Falcons.ai model scored highest in relevance, while Pegasus-XSum excelled in readability metrics like FKGL and LENS. The original extractive summaries outperformed others in factuality. The team also created a dictionary of medical terms translated to lay-terms for future use. Their work highlights both economic and practical accessibility in medical summarization.

**TLPIQ** (Bechler et al., 2025) This team focused on generating biomedical lay summaries using a fine-tuned FLAN-T5 base model, leveraging abstract and conclusion sections of articles along with expert-written lay summaries. They improved accessibility and understanding by maintaining factuality and domain relevance, despite falling short on readability compared to larger models like Llama3 and Qwen2.5. Their approach included instruction tuning with dataset tags and a specialized prompt template, achieving competitive relevance and superior factuality scores. However, the model's readability could be further enhanced through strategies such as dataset-specific training and post hoc lexical simplification.

**LaySummX** (Lin and Yu, 2025) This team introduced a retrieval-augmented fine-tuning framework for biomedical lay summarization, utilizing abstract-driven semantic retrieval with LoRA-tuned LLaMA3.1 models. By incorporating relevant full-text segments retrieved using the article abstracts into the fine-tuning process, they improved relevance and factuality metrics significantly compared to base models and individually tuned models, while maintaining competitive readability. Their method efficiently addresses computational constraints by segmenting articles into manageable units, demonstrating strong performance among open-source systems and closed-source models like GPT.

**5cNLP** (Lossio-Ventura et al., 2025) This team leveraged a combination of prompting strategies, retrieval techniques, and multimodal fusion for generating lay summaries from scientific articles and radiology reports. They utilized structured (compositional) prompting with role-based instructions to guide large language models (LLMs) like Llama-3.3-70B-Instruct and GPT-4.1 in producing summaries that are accessible to a general audience. Their method also incorporated retrieval-augmented generation (RAG) using biomedical knowledge from UMLS to enrich context understanding and employed a multimodal pipeline combining images and captions for radiology report summarization. Notably, their approach achieved second place in Subtask 2.1 close-source track and third place in Subtask 1.2, demonstrating the effectiveness of their framework in improving accessibility and understandability of complex medical information.

**MIRAGES** (Pong et al., 2025) The team approached the BioLaySumm 2025 task by building on an extract-then-summarize framework, emphasizing the importance of high-quality data curation for biomedical lay summarization. They experimented with various extractive summarization strategies and employed LoRA to fine-tune a Llama-3-8B to enhance readability and factual accuracy of downstream abstractive summaries. Additionally, they explored counterfactual data augmentation and post-processing definition insertion to further improve factual grounding and accessibility. Their system ranked 4th overall and achieved 2nd place in readability, demonstrating that good input design and targeted fine-tuning are critical for effective biomedical lay summarization. Their findings suggest that strategic data curation can have a more positive impact than merely increasing the

volume of fine-tuning samples in domain-specific summarization tasks.

**SUWMIT** (Basu et al., 2025) This team developed an open-source, end-to-end pipeline for the automated generation of lay summaries from biomedical articles, achieving top scores in two out of four relevance metrics and the highest overall ranking in the plain lay summarization subtask. Their approach involved fine-tuning a Llama-3.1-8B model with LoRA, utilizing a contrastive decoding strategy known as DoLa to improve factuality and readability. They experimented with various preprocessing, extractive summarization, and abstractive summarization techniques, ultimately finding that including Flesch-Kincaid grade-level targets in system messages and applying LoRA weights during decoding were crucial for their success. Additionally, they explored different data transformation methods, including the use of BioBERT embeddings for extractive summarization, to enrich input context for improved summary quality.

**KHU_LDI** (Moriazi and Sung, 2025) This team explored two approaches for generating lay radiology reports: supervised fine-tuning of open-source large language models using QLoRA, and a refinement process to improve the initial generated output. They found that while the fine-tuned model outperformed the refinement approach on test data, the refinement method showed significant potential on the validation set, particularly when using GPT-4o-mini as both the feedback and refinement models. Their submission achieved second place in the open track of Subtask 2.1, highlighting the effectiveness of fine-tuning open-source models for producing patient-friendly radiology reports.

**BDA-UC3M** (Ramzi and Bedmar, 2025) This team focused on demonstrating that small-scale, state-of-the-art language models (4B–7B parameters) can achieve competitive performance in biomedical lay summarization. Utilizing models such as Gemma3 4B, Qwen3 4B, and GPT-4.1-mini, they employed dynamic 4-bit quantization, extractive preprocessing, prompt engineering, data augmentation, and Direct Preference Optimization to enhance efficiency and factuality. Their system ranked second in its category by generating high-quality, accurate summaries, highlighting the potential of compact models for making complex scientific content accessible to non-expert audiences without sacrificing performance.

**CUTN_Bio** (Sivagnanam et al., 2025) This team focused on developing a prompt-based lay summarization framework for biomedical articles and radiology reports as part of the BioLaySumm 2025 shared task. For plain lay summarization, they utilized Llama-3-8B with a Tree-of-Thought prompting strategy to generate simplified summaries. In the lay summarization with external knowledge subtask, they combined an extractive approach (LEAD-K sentence extraction) with Llama-3-8B, enriched by medical definitions from MedCAT and Wikipedia, achieving the second position in Task 2.1. For radiology report translation, they implemented a Retrieval-Augmented Generation (RAG) method using the Zephyr model, achieving third in this category. Their methodologies highlight the effectiveness of combining external knowledge, extractive summarization techniques, and instruction-tuned language models for generating accessible summaries.

## 6 Results Analysis

The BioLaySumm 2025 shared task revealed critical insights about biomedical lay summarization methodologies, emphasizing trade-offs, architectural innovations, and emerging trends. The analysis below synthesizes key findings from both the competition leaderboard (Table 2) and participant approaches.

**Trade-offs Between Evaluation Metrics** No single system dominated all evaluation dimensions (relevance, readability, factuality), revealing inherent conflicts in optimization objectives. For instance, SUWMT (1st in Subtask 1.1) excelled in relevance (ROUGE: 0.370) but produced complex text (FKGL: 11.74), while MetinOZU achieved exceptional factuality (SummaC: 0.920) at the cost of poor readability (FKGL: 16.45). Aard demonstrated balanced readability (FKGL: 11.16) and factuality (SummaC: 0.537) in Subtask 1.2 but lagged in relevance (ROUGE: 0.292). These cases illustrate how excelling in one metric often compromises others, necessitating task-specific customization.

**Dominance of Retrieval-Augmented Generation** Retrieval-augmented approaches emerged as a dominant trend, with 5 of 13 teams (LaySummX, BioSumEnhance, CUTN_Bio, Aard, and 5cNLP) incorporating external knowledge. This strategy

proved particularly effective in Subtask 1.2 (external knowledge), where Aard and CUTN_Bio secured 1st and 2nd places with 7–9% factuality gains over non-RAG baselines. Teams leveraged UMLS, Wikipedia, and full-text segments to handle domain terminology, though sometimes at the cost of readability due to verbose outputs.

**Persistence of Pipeline Approaches** Pipeline frameworks remained prevalent, with 7 of 13 teams adopting multi-stage architectures rather than unified models. Examples include MIRAGES' extract-then-summarize approach using extractive summarization followed by LoRA-tuned Llama3-8B (ranking 6th with 2nd-best readability), and Aard's modular system combining BioBERT-based entity recognition with iterative rewriting. These pipelines offered interpretability advantages but introduced potential error propagation risks compared to end-to-end systems like SUWMT's top-ranked submission.

**Competitiveness of Legacy Architectures** Encoder-decoder models demonstrated comparable performance against larger LLMs. AEHRC's T5-large (700M parameters) outperformed 3B+ LLMs in 9 of 10 metrics for radiology report translation (Subtask 2.1), dominating both competition tracks. Similarly, TLPIQ's FLAN-T5 base model achieved competitive relevance and factuality despite its smaller size, underscoring the continued efficiency of these architectures for domain-specific generation tasks.

**Emerging Methodological Innovations** Several novel techniques showed promise: XSZ explored reinforcement learning (PPO/GRPO) with multi-objective rewards optimizing factuality, readability, and relevance; MetinOZU developed reverse data augmentation by generating complex scientific text from simple summaries; and BDA-UC3M implemented efficiency techniques like 4-bit quantization with Direct Preference Optimization. While not all innovations were competition submissions, they represent significant research directions.

**Hardware Efficiency Demonstrations** Several teams validated cost-effective approaches, most notably RainCityNLP which combined TF-IDF sentence scoring with Pegasus-XSum and medical Falcons.ai models running on consumer-grade hardware. These implementations demonstrate the feasibility of deploying lay summarization systems in resource-constrained environments while maintaining reasonable performance.

**Key Gaps and Future Directions** Three critical challenges emerged from the analysis: (1) The persistent conflicts between readability and factuality require new joint optimization strategies; (2) External knowledge integration through RAG sometimes disrupted narrative coherence despite improving accuracy; (3) Reinforcement learning approaches like XSZ's show untapped potential for metric-aligned reward shaping that warrants deeper exploration.

## 7 Conclusion

The third edition of the BioLaySumm Shared Task was hosted by the BioNLP Workshop@ACL 2025. Several changes were implemented over the previous edition, including the incorporation of the new task, lay summarization of radiology reports. The competition outcomes underscore biomedical lay summarization as a multi-faceted challenge requiring context-aware solutions. While RAG and pipeline methods dominated submissions, legacy encoder-decoder models (T5, FLAN-T5) remained surprisingly effective. Future work should prioritize hybrid approaches, particularly RAG-enhanced end-to-end models with RL fine-tuning, to better harmonize the competing demands of relevance, readability, and factuality.

## References


21st Century Cures Act. 2016. 21st century cures act. An Act to accelerate the discovery, development, and delivery of 21st century cures, and for other purposes.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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

Priyam Basu, Jose Cols, Daniel Jarvis, Yongsin Park, and Daniel Rodabaugh. 2025. Suwmit at biolaysumm2025: Instruction-based summarization with contrastive decoding. In *The 24th Workshop*

*on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Melody Bechler, Carly Crowther, Emily Luedke, Natasha Schimka, and Ibrahim Sharaf. 2025. Tlpiq at biolaysumm: Hide and seq, a flan-t5 model for biomedical summarization. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

Dina Demner-Fushman, Sameer Antani, Matthew Simpson, and George R Thoma. 2012. Design and development of a multimodal biomedical information retrieval system. *Journal of Computing Science and Engineering*, 6(2):168–177.

Egecan Çelik Evgin, lknur Karadeniz, and Olcay Taner Yıldız. 2025. Metninozu at biolaysumm2025: Text summarization with reverse data augmentation and injecting salient sentences. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chengua Lin. 2023. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 468–477. Association for Computational Linguistics.

Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024. Overview of the biolaysumm 2024 shared task on the lay summarization of biomedical research articles. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 122–131.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao,

Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Aaradhya Gupta and Dr Parameswari Krishnamurthy. 2025. Shared task at biolaysumm2025 : Extract then summarize approach augmented with umls based definition retrieval for lay summary generation. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

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.

Fan Lin and Dezhi Yu. 2025. Laysummx at biolaysumm: Retrieval-augmented fine-tuning for biomedical lay summarization using abstracts and retrieved full-text context. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Juan Antonio Lossio-Ventura, Callum Chan, Arshitha Basavaraj, Hugo Alatrista-Salas, Francisco Pereira, and Diana Inkpen. 2025. 5cnlp at biolaysumm2025: Prompts, retrieval, and multimodal fusion. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

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

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.

Teresa Martin-Carreras, Tessa S Cook, and Charles E Kahn Jr. 2019. Readability of radiology reports: implications for patient-centered care. *Clinical imaging*, 54:116–120.

Nur Alya Dania binti Moriazi and Mujeen Sung. 2025. Khu_ldi at biolaysumm2025: Fine-tuning and refinement for lay radiology report generation. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

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

Catia Pesquita, Daniel Faria, Andre O Falcao, Phillip Lord, and Francisco M Couto. 2009. Semantic similarity in biomedical ontologies. *PLoS computational biology*, 5(7):e1000443.

Benhamin Pong, Ju-Hui Chen, Jonathan Jiang, Abimael Hernandez Jimenez, and Melody Vahadi. 2025. Mirages at biolaysumm2025: The impact of search terms and data curation for biomedical lay summarization. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Ilyass Ramzi and Isabel Segura Bedmar. 2025. Bdauc3m @ biolaysumm: Efficient lay summarization with small-scale sota llms. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Bhuvaneswari Sivagnanam, Rivo Krishnu C H, Princi Chauhan, and Saranya Rajiakodi. 2025. Cutn_bio at biolaysumm: Multi-task prompt tuning with external knowledge and readability adaptation for layman summarization. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. 2020. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*.

Jen Wilson, Avery Bellamy, Rachel Edwards, Michael Pollack, and Helen Salgi. 2025. Raincitynlp at biolaysumm2025: Extract then summarize at home. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Pengcheng Xu, Sicheng Shen, Jieli Zhou, and Hongyi Xin. 2025. Team xsz at biolaysumm2025: Section-wise summarization, retrieval-augmented llm, and reinforcement learning fine-tuning for lay summaries. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. 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.

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.

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

Wenjun Zhang, Shekhar S. Chandra, Bevan Koopman, Jason Dowling, and Aaron Nicolson. 2025. Aehrc at biolaysumm 2025: Leveraging t5 for lay summarisation of radiology reports. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Kun Zhao, Chenghao Xiao, Sixing Yan, Haoteng Tang, William K. Cheung, Noura Al Moubayed, Liang Zhan, and Chenghua Lin. 2025. X-ray made simple: Lay radiology report generation and robust evaluation.

# A Appendix

Table 3 and Table 4 present the overview and detailed metric performance after min-max normalization.

| # | Team | Relevance | Readability | Factuality | Avg. |
|---|---|---|---|---|---|
| 1 | **SUWMIT** | **0.971** | 0.816 | 0.616 | **0.801** |
| 2 | **Baseline-llama3-8B-sft** | 0.965 | 0.770 | 0.548 | 0.761 |
| 3 | **Baseline-qwen2.5-7B-sft** | 0.922 | 0.645 | 0.579 | 0.715 |
| 4 | **BDA-UC3M** | 0.892 | 0.726 | 0.406 | 0.675 |
| 5 | **MetninOzU** | 0.779 | 0.148 | **1.000** | 0.643 |
| 6 | **MIRAGES** | 0.531 | 0.886 | 0.466 | 0.628 |
| 7 | **TLP1Q** | 0.775 | 0.505 | 0.586 | 0.622 |
| 8 | **LaySummX** | 0.663 | 0.752 | 0.371 | 0.595 |
| 9 | **CUTN_Bio** | 0.470 | **0.972** | 0.316 | 0.586 |
| 10 | **Aard** | 0.708 | 0.505 | 0.376 | 0.530 |
| 11 | **LTRC** | 0.522 | 0.725 | 0.250 | 0.499 |
| 12 | **5cNLP** | 0.732 | 0.432 | 0.256 | 0.473 |
| 13 | **RainCityNLP** | 0.525 | 0.056 | 0.451 | 0.344 |
| 14 | **sxz** | 0.004 | 0.334 | 0.560 | 0.299 |
| 15 | **demo** | 0.004 | 0.334 | 0.560 | 0.299 |
| 16 | **x2z** | 0.054 | 0.787 | 0.014 | 0.285 |

(a) Subtask 1.1: Plain Lay Summarization

| # | Team | Relevance | Readability | Factuality | Avg. |
|---|---|---|---|---|---|
| 1 | **Aard** | 0.696 | **1.000** | **0.879** | **0.858** |
| 2 | **CUTN_Bio** | 0.667 | 0.382 | 0.861 | 0.637 |
| 3 | **5cNLP** | **1.000** | 0.039 | 0.447 | 0.495 |
| 4 | **LTRC** | 0.000 | 0.327 | 0.000 | 0.109 |

(b) Subtask 1.2: Lay Summarization with External Knowledge

| # | Team | Relevance | Readability | Clinical | Avg. |
|---|---|---|---|---|---|
| 1 | **AEHRC** | **1.000** | **0.667** | **1.000** | **0.889** |
| 2 | **KHU_LDI** | 0.000 | 0.333 | 0.000 | 0.111 |

(c) Subtask 2.1: Radiology Report Translation (Open Track)

| # | Team | Relevance | Readability | Clinical | Avg. |
|---|---|---|---|---|---|
| 1 | **AEHRC** | **1.000** | 0.521 | **0.858** | **0.793** |
| 2 | **Baseline-qwen2.5-7B-sft** | 0.567 | 0.384 | 0.601 | 0.517 |
| 3 | **5cNLP** | 0.688 | 0.300 | 0.537 | 0.508 |
| 4 | **Baseline-llama3-8B-sft** | 0.510 | 0.262 | 0.718 | 0.497 |
| 5 | **CUTN_Bio** | 0.000 | **0.809** | 0.000 | 0.270 |

(d) Subtask 2.1: Radiology Report Translation (Closed Track)

Table 3: Task leaderboard with min–max normalization

| Rank | Team | Relevance | | | | Readability | | | | Factuality | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ROUGE | BLEU | MTR | BERTS | FKGL | DCRS | CLI | LENS | AlignS | SummaC |
| 1 | **SUWMIT** | **1.000** | **1.0000** | 0.964 | 0.921 | 0.804 | 0.815 | 0.793 | 0.851 | 0.745 | 0.487 |
| 2 | **Baseline-llama3-8B-sft** | 0.977 | 0.9770 | **1.000** | 0.906 | 0.730 | 0.773 | 0.721 | 0.855 | 0.690 | 0.405 |
| 3 | **Baseline-qwen2.5-7B-sft** | 0.909 | 0.8504 | 0.933 | 0.996 | 0.648 | 0.646 | 0.593 | 0.692 | 0.753 | 0.406 |
| 4 | **BDA-UCM** | 0.917 | 0.7759 | 0.876 | **1.000** | 0.710 | 0.763 | 0.688 | 0.742 | 0.631 | 0.181 |
| 5 | **MetinOZU** | 0.880 | 0.6490 | 0.854 | 0.812 | 0.046 | 0.183 | 0.000 | 0.365 | **1.000** | **1.000** |
| 6 | **MIRAGES** | 0.598 | 0.3886 | 0.482 | 0.656 | 0.809 | **1.000** | 0.900 | 0.834 | 0.611 | 0.322 |
| 7 | **TupiQ** | 0.829 | 0.6733 | 0.717 | 0.862 | 0.531 | 0.369 | 0.642 | 0.478 | 0.762 | 0.642 |
| 8 | **LaySummX** | 0.759 | 0.4798 | 0.623 | 0.793 | 0.718 | 0.690 | 0.650 | 0.953 | 0.600 | 0.142 |
| 9 | **CUTN_Bio** | 0.503 | 0.2329 | 0.457 | 0.690 | **1.000** | 0.888 | **1.000** | **1.000** | 0.431 | 0.202 |
| 10 | **Aard** | 0.749 | 0.4805 | 0.871 | 0.730 | 0.350 | 0.537 | 0.295 | 0.837 | 0.637 | 0.118 |
| 11 | **LTRC** | 0.599 | 0.3473 | 0.430 | 0.711 | 0.543 | 0.752 | 0.667 | 0.938 | 0.455 | 0.045 |
| 12 | **5cNLP** | 0.820 | 0.5577 | 0.713 | 0.838 | 0.108 | 0.425 | 0.300 | 0.896 | 0.513 | 0.000 |
| 13 | **RainCityNLP** | 0.582 | 0.4152 | 0.544 | 0.561 | 0.000 | 0.051 | 0.138 | 0.037 | 0.476 | 0.426 |
| 14 | **sxz** | 0.000 | 0.0169 | 0.000 | 0.000 | 0.668 | 0.000 | 0.669 | 0.000 | 0.964 | 0.157 |
| 15 | **demo** | 0.000 | 0.0169 | 0.000 | 0.000 | 0.668 | 0.000 | 0.669 | 0.000 | 0.964 | 0.157 |
| 16 | **x2z** | 0.085 | 0.0000 | 0.094 | 0.036 | 0.666 | 0.713 | 0.782 | 0.730 | 0.000 | 0.028 |

(a) SubTask 1.1: Plain Lay Summarization

| Rank | Team | Relevance | | | | Readability | | | | Factuality | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ROUGE | BLEU | MTR | BERTS | FKGL | DCRS | CLI | LENS | AlignS | SummaC |
| 1 | **Aard** | 0.643 | 0.592 | 0.882 | 0.665 | **1.000** | **1.000** | **1.000** | **1.000** | 0.757 | **1.000** |
| 2 | **CUTN_Bio** | 0.676 | 0.532 | 0.559 | 0.902 | 0.570 | 0.020 | 0.150 | 0.787 | **1.000** | 0.722 |
| 3 | **5cNLP** | **1.000** | **1.000** | **1.000** | **1.000** | 0.000 | 0.000 | 0.000 | 0.155 | 0.745 | 0.149 |
| 4 | **LTRC** | 0.000 | 0.000 | 0.000 | 0.000 | 0.487 | 0.328 | 0.495 | 0.000 | 0.000 | 0.000 |

(b) Subtask 1.2: Lay Summarization with External Knowledge

| Rank | Team | Relevance | | | | | Readability | | | Clinical Metrics | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ROUGE | BLEU | MTR | BERTS | SIM | FKGL | DCRS | CLI | CHEX | RG |
| 1 | **AEHRC** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.000 | **1.000** | **1.000** | **1.000** |
| 2 | **KHU_LDI** | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | **1.000** | 0.000 | 0.000 | 0.000 |

(c) Subtask 2.1: Radiology Report Translation (Open Track)

| Rank | Team | Relevance | | | | | Readability | | | Clinical Metrics | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ROUGE | BLEU | MTR | BERTS | SIM | FKGL | DCRS | CLI | CHEX | RG |
| 1 | **AEHRC** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | 0.294 | 0.72 | 0.548 | 0.715 | **1.000** |
| 2 | **Baseline-qwen2.5-7B-sft** | 0.591 | 0.449 | 0.483 | 0.721 | 0.589 | **1.000** | 0.06 | 0.094 | 0.733 | 0.468 |
| 3 | **5cNLP** | 0.670 | 0.555 | 0.760 | 0.685 | 0.770 | 0.000 | 0.000 | 0.557 | 0.446 | 0.627 |
| 4 | **Baseline-llama3-8B-sft** | 0.546 | 0.427 | 0.414 | 0.649 | 0.512 | 0.786 | 0.000 | 0.000 | **1.000** | 0.436 |
| 5 | **CUTN_Bio** | 0.404 | 0.427 | 0.000 | 0.000 | 0.000 | 0.428 | **1.000** | **1.000** | 0.000 | 0.000 |

(d) Subtask 2.1: Radiology Report Translation (Closed Track)

Table 4: Task leaderboard with min-max normalization. **BertS** = BertScore, **FKGL** = Flesch-Kincaid Grade Level, **DCRS** = Dale-Chall Readability Score, **CLI** = Coleman-Liau Index, **MTR** = METOR, **SIM** = Similarity, **AlignS** = AlignScore, **CHEX** = F1 chexbert, **RG** = Radgraph.