AEHRC at BioLaySumm 2025: Leveraging T5 for Lay Summarisation of Radiology Reports

Wenjun Zhang[⋆†] **Shekhar S. Chandra**[⋆] **Bevan Koopman**[†]
Jason Dowling[†] **Aaron Nicolson**[†]
[⋆]The University of Queensland, Brisbane, Australia
[†]Australian e-Health Research Centre, CSIRO Health and Biosecurity, Brisbane, Australia
`wenjun.zhang@uq.edu.au`

Abstract

Biomedical texts, such as research articles and clinical reports, are often written in highly technical language, making them difficult for patients and the general public to understand. The **BioLaySumm 2025 Shared Task** addresses this challenge by promoting the development of models that generate lay summarisations of biomedical content. This paper focuses on Subtask 2.1: Radiology Report Generation with Layman's Terms. In this work, we evaluate two large language model (LLM) architectures, T5-large (700M parameter encoder–decoder model) and LLaMA-3.2-3B (3B parameter decoder-only model). Both models are trained under fully-supervised conditions using the task's multi-source dataset. Our results show that T5-large consistently outperforms LLaMA-3.2-3B across nine out of ten metrics, including relevance, readability, and clinical accuracy, despite having only a quarter of the parameters. **Our T5-based model achieved the top rank in both the open-source and close-source tracks of the subtask 2.1.**

1 Introduction

Biomedical texts, ranging from research articles to clinical reports, are often written in highly technical language. This presents a major barrier for patients and the general public, limiting their ability to understand health information and make informed decisions. As a response, the field of biomedical lay summarisation has emerged to simplify expert-level content into language that non-experts can understand (Guo et al., 2021). A particular application of this is the translation of radiology reports. A recent analysis found that only about 4% (Martin-Carreras et al., 2019) of radiology reports are written below an 8th-grade reading level (the average adult reading level). Without clear explanations, the information can be confusing or even misleading, creating barriers to understanding one's own health.

To advance research in this area, the BioLaySumm 2025 Shared Task hosted at the BioNLP Workshop at ACL 2025, introduces two primary tasks aimed at improving the accessibility of biomedical information for non-expert audiences (Xiao et al., 2025). Task 1 focuses on the lay summarisation of biomedical research articles, and Task 2 centers on lay summarisation of radiology reports. Participants are provided with a dataset containing paired radiology reports and lay summarisations from sources such as Open-i (Demner-Fushman et al., 2016), PadChest (Bustos et al., 2020), BIMCV-COVID19 (de la Iglesia Vayá et al., 2020), and MIMIC-CXR (Johnson et al., 2019). The shared task includes two evaluation tracks. The open-source track uses test data from three public datasets (Open-i, PadChest, and BIMCV-COVID19), while the close-source track adds the additional MIMIC-CXR dataset. In this work, we trained our models exclusively on the open-source datasets and submitted predictions for both the open-source and close-source tracks.

While language models have become central to biomedical lay summarisation (Xie et al., 2023), there remains no clear consensus on whether encoder-decoder or decoder-only architectures are better suited to this task. Prior studies have shown both to be effective in different contexts, but few have directly compared them under controlled, fully-supervised conditions. Motivated by this, we conduct a comparison between an encoder–decoder model T5-large (Raffel et al., 2020) and a decoder-only LLaMA-3.2-3B model (Meta AI, 2024). Our final submission, based on T5-large, achieved the highest overall score in Subtask 2.1, ranking first among all participating teams. It consistently outperforms LLaMA-3.2-3B across nine of ten evaluation metrics, despite having only a quarter of the parameters.

BioNLP 2025 Shared Tasks, pages 171–178
August 1, 2025 ©2025 Association for Computational Linguistics

2 Related Work

2.1 Biomedical Lay Summarisation

Biomedical lay summarisation is a sequence-to-sequence task that translates specialised medical language into accessible explanations for non-experts, such as patients (Xie et al., 2023). While biomedical summarisation has traditionally focused on preserving factual accuracy and completeness, much less attention has been given to simplifying language for broader public understanding. In response to this need, the BioLaySumm Shared Task series (2023–present) was introduced to encourage the development of models that generate lay summaries from biomedical content (Goldsack et al., 2023, 2024; Xiao et al., 2025).

Over time, the field of biomedical summarisation has evolved from rule-based and statistical methods to neural architectures and, more recently to language models (Xie et al., 2023). A similar trend is evident in the BioLaySumm Shared Tasks: in 2023, only 3 teams used LLMs, while in 2024, 18 teams used decoder-only LLMs (BioGPT (Luo et al., 2022), BioMistral (Labrak et al., 2024), GPT-4 (OpenAI, 2023), LLaMA (Touvron et al., 2023)) and 13 teams using encoder-decoder LLMs (T5 (Raffel et al., 2020), FLAN-T5 (Chung et al., 2024)), reflecting a growing interest in applying both architectures to the task of lay summarisation (Goldsack et al., 2024).

2.2 Encoder-Decoder and Decoder-only Language Models

Both encoder–decoder and decoder-only LLMs are based on the Transformer architecture (Vaswani et al., 2017), which was originally introduced for sequence-to-sequence tasks such as machine translation. The Transformer consists of two components: an encoder that processes the input into a latent space, and a decoder that generates the output based on that representation and the previously generated tokens. Encoder–decoder models, such as T5 (Raffel et al., 2020) and BART (Lewis et al., 2020), leverage both components to perform a wide range of text-to-text tasks, including translation, summarization, and question answering. In contrast, decoder-only models, such as the GPT series (OpenAI, 2022, 2023) and LLaMA (Touvron et al., 2023), comprise only the decoder component. These models are trained using an autoregressive objective, predicting the next token based solely on preceding tokens in a unidirectional manner. In the context of biomedical summarisation, both encoder–decoder and decoder-only models remain active areas of research, with no definitive answer as to which performs better (Goldsack et al., 2024). In this work, we will compare an encoder-decoder model T5 to a decoder-only model LLaMA on the task of lay summarisation of radiology reports.

2.3 Lay Summarisation of Radiology Reports

Due to the lack of alignment data, initial studies investigated the feasibility of direct prompting LLMs such as ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023), to simplify radiology reports. The generated contents are reviewed by radiologists to measure accuracy, completeness, and simplicity. For instance, lyu (2023) showed that while ChatGPT could generate understandable plain-language summaries, it often over-simplified or introduced inaccuracies. Similarly, Kuckelman et al. (2024) found that ChatGPT's summaries of musculoskeletal radiology reports were generally accurate but sometimes lacked completeness. In 2024, Zhao et al. (2025) introduced the LaymanRRG framework, utilised GPT-4o to translate professional radiology reports from MIMIC-CXR into layman's terms. BioLaySumm Shared Task 2 (Xiao et al., 2025) built on this dataset to establish a standardised benchmark for evaluating lay summarisation systems across four datasets Open-i, PadChest, and BIMCV-COVID19 and MIMIC-CXR (Demner-Fushman et al., 2016; Bustos et al., 2020; de la Iglesia Vayá et al., 2020; Johnson et al., 2019).

3 Methodology

Our primary goal was to develop a model for Subtask 2.1 of the BioLaySumm 2025 Shared Task, which involved generating lay summarisations from radiology reports. As part of this effort, we fine-tuned T5-large (Raffel et al., 2020) and LLaMA-3.2-3B (Meta AI, 2024) on the shared task dataset. The model with the highest validation BERTScore-F1 (Zhang et al., 2020) was submitted for testing on both the open-source track and the close-source track, described in Section 3.2.

3.1 Model Selection

We selected two prominent open-source LLM architectures:

- **T5-large (Raffel et al., 2020)**: An encoder–decoder model pretrained on Colossal Clean Crawled Corpus (C4) dataset (comprises 156 billion tokens) (Dodge et al., 2021).

- **LLaMA-3.2-3B (Meta AI, 2024)**: A decoder-only model pretrained on up to 9 trillion tokens from publicly available sources, including Common Crawl, Wikipedia, and other web-based corpora.

3.2 Datasets

Task 2.1 includes two evaluation tracks. The open-source track evaluates models on test sets from three publicly available datasets: Open-i, PadChest, and BIMCV-COVID19 (Demner-Fushman et al., 2016; Bustos et al., 2020; de la Iglesia Vayá et al., 2020). The close-source track adds an additional dataset, namely MIMIC-CXR (Johnson et al., 2019). We trained our models exclusively on the three open datasets, and submitted the same model to both the open-source track and close-source track. This setup enables us to evaluate in-domain performance (on seen datasets) as well as generalisation to unseen data (MIMIC-CXR). Table 1 summarises the dataset composition across tracks.

Table 1: Composition of datasets for Subtask 2.1. We train on the three open-source track datasets (top). The MIMIC-CXR dataset (bottom) is used exclusively for the close-source track and is not included in our training or validation.

Dataset	**Train**	**Val**	**Test**
Open-source track datasets (training)			
Open-i	2,243	134	186
PadChest	116,847	7,824	7,130
BIMCV-COVID19	31,364	2,042	3,221
Open-source total	**150,454**	**10,000**	**10,537**
Close-source track addition (evaluation only)			
MIMIC-CXR	45,000*	5,000*	500
Close-source test total	–	–	**11,037**

* Provided for reference only; not used in training or validation.

3.3 Data Processing and Tokenisation

For training T5, input and output sequences were tokenised separately using a SentencePiece tokeniser. The input (e.g., a radiology report) was tokenised without an end-of-sequence token, while the output (e.g., a lay summarisation) includes the <eos> token at the end. For example, the input “The chest X-ray shows pneumonia.” was tokenised as `['_The', '_chest', ..., '_pneumonia.']`, and the output “There is an infection in the lungs.” was tokenised as `['_There', '_is', ..., '_lungs.', '<eos>']`. In contrast, LLaMA-3.2-3B used a byte-level BPE tokeniser and expected the input and output to be concatenated into a single sequence for decoder-only training. We used a special token <s> to separate the lay summarisation from the radiology report. Specifically, the input before tokenisation was “The chest X-ray shows pneumonia.<s>There is an infection in the lungs.</s>”, where </s> is the end-of-sentence token.

3.4 Training

We trained both models for six epochs using PyTorch Distributed Data Parallel (DDP) on a single H100 compute node with four GPUs. Each GPU processed a local batch size of 2, resulting in a global batch size of 8. We used the AdamW optimiser (Loshchilov and Hutter, 2019) with a learning rate of 2e-5 and a constant learning rate schedule with 500 warmup steps. Input sequences were truncated or padded to a maximum length of 2048 tokens, and during inference, the models were allowed to generate up to 300 tokens. BERTScore-F1 (Zhang et al., 2020) was used as the primary metric for model selection.

3.5 Metrics

The evaluation is based on three key categories defined by the BioLaySumm 2025 Shared Task: relevance, readability, and clinical accuracy.

Relevance measures how well the generated summaries align with the reference texts. ROUGE-1/2/L (Lin, 2004) are computed for lexical overlap; BLEU (Papineni et al., 2002) for 1-to-4-gram precision with brevity penalty; METEOR (Banerjee and Lavie, 2005) for word-level matches and ordering; BERTScore (Zhang et al., 2020) for embedding-based semantic similarity. Semantic similarity between radiology report and lay summarisations is also measure.

Readability evaluates how easy the generated text is to understand for a general audience. The Flesch–Kincaid Grade Level (FKGL) (Kincaid et al., 1975), Dale–Chall Readability Score (DCRS) (Dale and Chall, 1948), and the Coleman–Liau Index (CLI) (Coleman and Liau, 1975) are reported; lower values indicate simpler language.

Clinical accuracy assesses the factual correctness of medical content. This is captured by

Table 2: Performance of LLaMA-3.2-3B and T5-large across relevance, readability, and clinical-accuracy metrics (open-source track). Boldface indicates best score.

Model	ROUGE	BLEU	MET.	BERT.	FKGL†	DCRS†	CLI†	Sim.	RadF1	ChexF1
LLaMA-3.2-3B	64.20	42.90	67.70	94.90	7.46	9.38	**8.04**	87.90	37.20	85.00
T5-large-770M	**66.9**	**45.6**	**70.2**	**95.3**	**7.43**	**9.34**	8.11	**88.9**	**40.0**	**85.6**

† Lower values indicate better readability.

CheXbert-F1 (Smit et al., 2020), which gauges agreement on clinical findings, and RadGraph-F1 (Jain et al., 2021), which measures overlap of medical entities and their relations.

4 Results and Discussion

4.1 Comparing Encoder–Decoder Models and Decoder-only Models Results

We present our submission results for the open-source track in Table 2, comparing our fine-tuned models: T5-large and LLaMA-3.2-3B. T5-large outperforms LLaMA-3.2-3B across 9 out of 10 metrics, including all relevance (ROUGE, BLEU, METEOR, BERTScore, and semantic similarity) and clinical accuracy metrics (RadGraph-F1 and CheXbert-F1). Readability scores are similar across models, but T5 achieves lower Dale–Chall and FKGL scores, indicating slightly simpler vocabulary.

Notably, T5-large achieves these results with only one-quarter the parameters of LLaMA-3.2-3B, suggesting that encoder–decoder architectures may be more efficient for the supervised lay summarization task. Several factors may explain this performance difference. First, the encoder–decoder structure of T5 is inductively biased toward input–output transformation tasks such as summarisation and paraphrasing, potentially making it more effective in low-resource, domain-specific settings. Second, the larger capacity of LLaMA-3.2-3B may require more training data to fully optimise, and in limited-data scenarios, its performance could suffer due to underfitting or instability. Indeed, we observed more fluctuation in validation scores during LLaMA training, whereas T5's training curve was smoother and more consistent.

These quantitative results are further supported by qualitative case studies in Appendix A and B, where T5's outputs consistently demonstrate higher fidelity to the original clinical findings while offering clearer and more accessible lay-language paraphrasing. For example, in Case 1, T5 translates technical terms like "atelectasis and consolidation" into "lung collapse and solid areas," while LLaMA retains the original jargon, making the output less readable for general audiences. In Case 2, T5 maintains more specific references such as “heart problems related to the coronary arteries” and “chronic obstructive pulmonary disease (COPD),” whereas LLaMA tends to generalise or simplify more loosely.

Table 3: Performance of T5-large on the open-source track evaluation and close-source track evaluation. Boldface indicates best score.

Metric	Open	Close
ROUGE-L	**66.90**	58.66
BLEU	**45.61**	32.08
METEOR	**70.17**	62.68
BERTScore	**95.30**	94.33
FKGL†	**7.43**	7.65
DCRS†	9.34	**8.60**
CLI†	8.11	**7.88**
Semantic Score	88.88	**89.52**
RadGraph-F1	**39.96**	34.81
CheXbert-F1	**85.64**	68.20

4.2 Comparing Open-source Track and Close-source Track Submission Results

We submitted our model to both the open-source track and close-source track for evaluation, with results shown in Table 3. The close-source track evaluates model performance on a broader test set than the open-source track. Specifically, the open-source track includes test samples from three public datasets—Open-i, PadChest, and BIMCV-COVID19—while the close-source track adds MIMIC-CXR.

Our model was trained exclusively on the open-source track training set, which means that its performance on the close-source track reflects both in-domain evaluation (on seen sources) and out-of-domain generalisation (on unseen MIMIC-CXR examples). We can observe overall trends: compared to the open-source track results, the close-source track metrics drop across nearly all categories, especially BLEU, ROUGE-L, and CheXbert-F1. However, because the closed test set is a mixture of all four datasets and aggregate scores are reported, we cannot isolate exact performance on MIMIC-CXR. Also, due to the absence of comparison with T5-large trained on the closed-source track, the relative performance drop remains unknown. Therefore, we cannot fully conclude on the model's generalisation to unseen data.

Nevertheless, our model achieved the top rank in both the open-source and close-source tracks, according to the competition leaderboard, outperforming all other submissions, including those trained on the full close-source dataset.

4.3 Conclusion

This study investigated the task of translating radiology reports into lay summarisation as part of Subtask 2.1 of the BioLaySumm 2025 Shared Task. We compare the performance of two large language model (LLM) architectures: the decoder-only LLaMA-3.2-3B and the encoder–decoder T5-large. Our results demonstrate that the T5-large model consistently outperforms LLaMA-3.2-3B across a broad range of metrics, including relevance, readability, and clinical accuracy, despite having significantly fewer parameters. This finding suggests that encoder–decoder models may be more effective and parameter-efficient for fully-supervised summarisation tasks, such as lay summarisation of radiology reports.

References

2023. Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: results, limitations, and potential. *Visual Computing for Industry, Biomedicine, and Art*, 6(1).

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL*, pages 65–72.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and María de la Iglesia-Vayá. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797.

Hyung Won Chung, Le Hou, Shayne Longpre, Jason Wei, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25:1–53.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–20, 28.

María 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, Marisa Caparrós, Germán González, and Jose María Salinas. 2020. BIMCV COVID-19+: A large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2).

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *In EMNLP*, pages 1286–1305.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua 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, Toronto, Canada. 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, Bangkok, Thailand. Association for Computational Linguistics.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *AAAI*, pages 160–168.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre J. Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *NeurIPS*.

Alistair E. W. 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*, 6:317.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Research Branch Report 8-75.

Ian J. Kuckelman, Karla Wetley, Paul Hyunsoo Yi, and Andrew Bailey Ross. 2024. Translating musculoskeletal radiology reports into patient-friendly summaries using ChatGPT-4. *Skeletal Radiology*, 53:1621–1624.

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

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 *ACL*, pages 7871–7880.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

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

Meta AI. 2024. Llama 3.2: Multilingual and multimodal foundation models. `https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md`. Model card, version 3.2 (release date: 2024-09-25).

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. `https://openai.com/blog/chatgpt`.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, Philadelphia, PA.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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. In *EMNLP*, pages 1500 – 1519.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Tomas Goldsack, Sophia Ananiadou, Noura Al Moubayed, Liang Zhan, William Cheung, and Chenghua Lin. 2025. Overview of the biolaysumm 2025 shared task on lay summarization of biomedical research articles and radiology reports. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Qianqian Xie, Zheheng Luo, Benyou Wang, and Sophia Ananiadou. 2023. A survey for biomedical text summarization: From pre-trained to large language models. *Preprint*, arXiv:2304.08763.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*.

Kun Zhao, Chenghao Xiao, Sixing Yan, William K. Cheung, Kai Ye, Noura Al Moubayed, Liang Zhan, and Chenghua Lin. 2025. X-ray made simple: Lay radiology report generation and robust evaluation. *Preprint*, arXiv:2406.17911.

A Case Study 1

Table 4: Comparison of T5 and LLaMA outputs for a radiology report indicating COVID-19-type pneumonic infiltrates. T5 better paraphrases technical terms like "atelectasis" into lay-friendly expressions like "lung collapse," improving accessibility. LLaMA preserves more clinical terminology, which may limit interpretability. T5 also mirrors the structure of the ground truth better, enhancing relevance and readability.

Radiology Report	**Ground Truth Lay Summary**
Bands of **atelectasis and consolidation** in the mid-lung fields, along with **confluent opacities** in the lower mid-lung fields, also showing bilateral involvement. Findings, considering the reported history, could correspond to **COVID-19-type pneumonic infiltrates**. Evaluate progression and correlate with clinical and laboratory data.	There are areas of **lung collapse and thickening** in the middle parts of the lungs, along with **large blurry spots** in the lower middle areas, affecting both lungs. These findings, considering the patient's history, could be related to **the type of lung inflammation seen in COVID-19**. Further evaluation of the progression and correlation with clinical and lab data is needed.
T5-large Output	**LLaMA-3.2-3B Output**
There are **bands of lung collapse and solid areas** in the middle parts of the lungs, along with **merged cloudiness** in the lower middle parts of the lungs, affecting both sides. These findings, considering the reported history, could be signs of **COVID-19-type lung inflammation**. It is important to evaluate the progression and compare with clinical and lab data.	The middle areas of the lungs show signs of **atelectasis and consolidation**, and the lower middle areas of the lungs have **merged opacities** on both sides. These findings, considering the patient's history, could be related to **COVID-19-like lung infiltrates**. It is recommended to evaluate the progression and correlate this with clinical and laboratory data.

B Case Study 2

Table 5: Comparison of T5 and LLaMA outputs for a radiology report covering dyspnea, coronary artery disease, scoliosis, and COPD. T5 demonstrates better consistency with the ground truth by retaining specific phrases like "coronary artery disease" and clearly expressing medical conditions. LLaMA tends to generalise phrases (e.g., "heart problems"), which may lose important details for patient understanding.

Radiology Report	**Ground Truth Lay Summary**
Reason for consultation: **dyspnea**. **Old coronary artery disease**. **Significant dorsal scoliosis**. **Aortic elongation with descending aortic dilatation**. **Right pleural thickening**. Radiological signs of **chronic obstructive pulmonary disease (COPD)**.	The person came in because they were having **trouble breathing**. They have a history of **heart problems related to the coronary arteries**. They also have a **significant curve in their spine**. Their **main blood vessel, the aorta, is longer than usual and the part of it that goes down from the heart is wider**. The **outer layer of the lung on the right side is thicker than it should be**. The x-ray shows signs of a **lung disease called COPD, which makes it hard to breathe**.
T5-large Output	**LLaMA-3.2-3B Output**
The patient came in because they were having **trouble breathing**. They have a history of **heart problems related to the coronary arteries**. They also have a **significant curve in their spine**. The **main blood vessel coming out of the heart is stretched out and wider than normal**. The **outer lining of the lungs on the right side is thicker than it should be**. The x-ray shows signs of a **long-term lung condition called chronic obstructive pulmonary disease (COPD)**.	The patient came in because they were having **trouble breathing**. They have a history of **heart problems** and a **curved spine**. The **main blood vessel leaving the heart, called the aorta, is stretched out and the part that goes down from the heart is widened**. The **outer lining of the lungs on the right side is thicker than normal**. The x-ray shows signs of a **long-term lung condition called COPD**.