KHU_LDI at BioLaySumm2025: Fine-tuning and Refinement for Lay Radiology Report Generation

Nur Alya Dania binti Moriazi and Mujeen Sung
Kyung Hee University, South Korea
{dania.moriazi01, mujeensung}@khu.ac.kr

Abstract

Though access to one's own radiology reports has improved over the years, the use of complex medical terms makes understanding these reports difficult. To tackle this issue, we explored two approaches: supervised fine-tuning open-source large language models using QLoRA, and refinement, which improves a given generated output using feedback generated by a feedback model. Despite the fine-tuned model outperforming refinement on the test data, refinement showed good results on the validation set, thus showing good potential in the generation of lay radiology reports. Our submission achieved 2nd place in the open track of Subtask 2.1 of the BioLaySumm 2025 shared task.

1 Introduction

There has been a growing demand in recent years for patients' ability to access their own medical records, particularly their radiology reports (Steitz et al., 2023, Vincoff et al., 2022). However, even when made accessible, radiology reports, as written by radiologists, are difficult to understand due to highly technical vocabulary. A 2019 review showed that the majority of radiology reports required at least college-level reading skills, with only 4.2% of radiology reports being readable at the 8th-grade reading level or below (Martin-Carreras et al., 2019). The BioLaySumm 2025 shared task addresses this issue by introducing a new task which aims to create patient-friendly (i.e. layman) versions of radiology reports (Xiao et al., 2025).

Large language models (LLMs) such as Qwen (Bai et al., 2023), LLaMA (Touvron et al., 2023) and GPT-4 (OpenAI et al., 2024b) have demonstrated notable ability in summarising medical texts (Das et al., 2025, Zhou et al., 2024). Likewise, the results of previous editions of the BioLaySumm shared task (Goldsack et al., 2023, Goldsack et al., 2024) have shown that LLMs are capable of producing lay versions of biomedical texts. Therefore, there is potential in using LLMs for the generation of lay radiology reports.

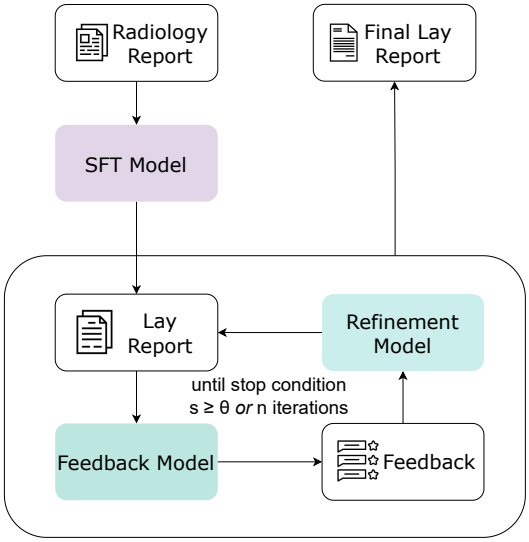

Figure 1: Our refinement framework for the lay radiology report generation task.

Recent research (Zhao et al., 2024, Sterling et al., 2024) has demonstrated the ability of OpenAI's GPT-3 (Brown et al., 2020) and GPT-4 models to generate lay radiology reports. However, OpenAI models can be costly over time thus potentially making lay radiology reports financially infeasible. As such, fine-tuning open-source LLMs may be more viable down the line for lay radiology report generation. Furthermore, fine-tuning allows open-source models to adapt to domain- or task-specific data. In the context of healthcare, this allows models to become familiar with medical vocabulary which, in turn, improves the quality of generated lay reports.

Welleck et al. (2022) and Madaan et al. (2023) have shown that, just as humans evaluate and edit

BioNLP 2025 Shared Tasks, pages 256–268

their own work, LLMs are not only capable of evaluating and refining their own outputs but also benefit from doing so. At the same time, lay reports must be readable and maintain factual accuracy. Despite the ability of LLMs to produce medical summaries, the results obtained are still riddled with hallucinations (Das et al., 2025). Therefore, we see refinement as a potential approach in ensuring readability whilst being faithful to the original professional lay reports.

We experimented with two approaches for the shared task: (1) supervised fine-tuning an open-source LLM and (2) refinement. We fine-tuned an LLM using QLoRA (Dettmers et al., 2023) on pairs of radiology reports and their corresponding lay reports to generate layman versions of radiology reports, and we used the GPT-4o-mini model (OpenAI et al., 2024a) to refine the output generated by the fine-tuned model. Although refinement showed promising results on the validation set, the model that performed best on the test data was a fine-tuned Qwen3-4B (Yang et al., 2025) model, which achieved 2nd place in the shared task.

2 Methods

2.1 Supervised Fine-Tuning (SFT)

We fine-tuned open-source large language models on pairs of radiology reports and their corresponding lay reports to train the model to generate a lay report given a professional radiology report. We performed SFT with QLoRA to optimise memory usage and increase efficiency. The prompt we used to fine-tune our models can be seen in Appendix A.

2.2 Refinement

We adapted the Self-Refine framework by Madaan et al. (2023) for lay radiology report generation (Figure 1). The refinement framework can be broken down into three steps: (1) Generation, (2) Feedback, and (3) Refinement.

Generation. We used an SFT model, M_{SFT}, for the initial generation. We used a few-shot prompt p_{gen} to generate the initial lay report y_0 given a professional radiology report x so that:

$$y_0 = M_{SFT}(p_{gen} \| x). \quad (1)$$

Feedback. Given a radiology report and generated lay report pair $< x, y_i >$, where $i \in n$ is the iteration step and n is the maximum number of iteration steps, feedback is generated for the lay report y_i using few-shot prompting on our feedback model, M_{fb}:

$$fb_i = M_{fb}(p_{fb} \| x \| y_i), \quad i = 0, 1, \ldots, n. \quad (2)$$

We generated a synthetic dataset containing radiology reports, generated lay reports and their feedback for our few-shot prompt p_{fb}. The lay reports used in our feedback dataset were generated by GPT-4o, and base open-source instruction models (particularly Llama-3.1-8B-instruct (Grattafiori et al., 2024) and Qwen2.5-7B-instruct (Yang et al., 2024)). We used the few-shot examples only on the first feedback p_{fb_0}.

To prevent M_{fb} from generating feedback that contradicts feedback from previous iterations, we appended previous feedback to p_{fb_i} where $1 \leq i \leq n$ in lieu of the examples from the feedback dataset for *iterative* refinement so,

$$fb_i = M_{fb}(p_{fb} \| x \| y_i \| \ldots \| y_0 \| fb_0). \quad (3)$$

As per Madaan et al. (2023), we prompted the model to encourage actionable feedback i.e. feedback that specifically pointed out sections of the text that should be improved on (see Figure 5 in Appendix A).

Refinement. We use a refinement model, M_R, to generate the refined lay report given the generated feedback, fb_i and radiology report-lay report pair $< x, y_i >$ so that

$$y_{i+1} = M_R(p_R \| x \| y_i \| fb_i). \quad (4)$$

Similar to the feedback step, we appended previous feedback and refined lay reports from previous iterations to the prompt for iterative refinement (see Figure 8 in Appendix A) to prevent M_R from generating outputs similar to previous iterations i.e. to learn from previous iterations so that

$$y_{i+1} = M_R(p_R \| x \| y_i \| fb_i \| \ldots \| y_0 \| fb_0). \quad (5)$$

Stop Condition. For iterative refinement, we employed a stop condition to control the number of iterations in the refinement framework. For this, we used two stop conditions: (1) a score threshold, $\theta = N_{\text{aspects}} \times 9$, where the score is extracted from fb_i, and (2) a set number of maximum iterations n. Refinement is performed iteratively until θ is reached or exceeded, or until n iterations are performed (whichever occurs first).

We detail our experiments with the refinement framework further in Section 3.3, where we discuss the different models used for feedback and refinement and the different aspects used by the feedback model to evaluate the lay reports.

3 Experiment Setup

The prompts we used for generation, feedback and refinement are detailed in Appendix A.

3.1 Data

We used the open-source track dataset provided by Xiao et al. (2025) for the second task of the BioLaySumm shared task, which is based on the LaymanRGG dataset by Zhao et al. (2024). The dataset comprises radiology images and their corresponding radiology reports and lay reports from the PadChest, BIMCV-COVID19 and OpenI datasets. Out of the three data sources, the PadChest dataset makes up the majority of the dataset, followed by the BIMCV-COVID19 and OpenI datasets (Table 1).

Source	Train	Validation	Test
PadChest	116,847	7,824	7,130
BIMCV-COVID19	31,364	2,042	3,221
OpenI	2,243	134	186
Total	150,454	10,000	10,537

Table 1: Number of samples from each data source in the dataset.

As we did not participate in the multi-modal version of the task, we did not use the radiology images in our experiments.

3.2 Supervised Fine-Tuning

We experimented with fine-tuning Qwen2.5-3B-Instruct and Qwen3-4B using QLoRA, which injects trainable low-rank adapter layers (LoRA) into specified model layers. We injected these layers into all the model's linear projection layers, as that tended to result in performance comparable to a fully fine-tuned model according to Dettmers et al. (2023).

We performed our experiments on an NVIDIA GeForce RTX 3090 graphics processing unit (GPU). We trained our models for 5 epochs with a learning rate of 5e-4 and an effective batch size of 128. For QLoRA, we set our rank $R = 64$ and $\alpha = 128$ to maximise performance whilst still training the model efficiently.

3.3 Refinement

We chose Qwen3-4b-SFT as our generation model as it showed the best performance on the validation set. For the feedback and refinement models, we experimented with using the SFT model for both feedback and refinement, using GPT-4o-mini (which performed best among the GPT models (see Table 2) on the validation set) for only feedback whilst using the SFT model for only refinement, and using GPT-4o-mini for both feedback and refinement (see Appendix B). Subsequently, we found that the framework that worked best was when we used GPT-4o-mini as both the feedback and the refinement models.

We initially had our feedback model evaluate the generated report on seven aspects: factuality, readability, completeness, conciseness, writing style (to avoid conversational language), format (to avoid verbose commentary), and structure (to discourage bullet points and lists). However, when examining the impact of each aspect on a single sample (see Appendix C), the aspects that showed significant improvement when used were completeness, factuality and format. Readability was shown to negatively impact the overall quality of the report, with improvements to the readability scores (section 4.1) being minimal compared to most of the other aspects.

We also experimented with iterative refinement on our validation sample set (see Section 4.2) with $n = 1, 3, 5$, where n is the number of iterations. Experiments show that a single iteration (i.e., without looping) consistently outperformed $n = 3$ and $n = 5$ when $max\_new\_token = 256$ for the first generation, but 3 iterations and 5 iterations consistently outperformed one iteration when $max\_new\_token = 512$ for the first generation, with $n = 3$ performing better than $n = 5$. Of the three iteration settings, the setting that performed the best was the 3-iteration setting with $max\_new\_token = 512$.

Furthermore, based on the scores extracted from the feedback, experiments conducted to evaluate the necessity of few-shot feedback prompting and the inclusion of past history found that few-shot feedback prompting on the first iteration and the inclusion of past history in subsequent iterations consistently resulted in an improvement of scores with each iteration across all model and iteration settings (provided that $n \neq 1$), whilst using only few-shot feedback prompting (in all iterations) or only including past history or using neither tended to result in a decrease in scores with each iteration.

These experiments found that the best refinement setting was $max\_new\_token = 512$, $n = 3$ with few-shot feedback prompting on the first iteration and the inclusion of past history.

Model	Relevance					Readability			Clinical		Average
	ROUGE	BLEU	METEOR	BERTScore	Semantic	FKGL	DCRS	CLI	F1CheXbert	F1RadGraph	
GPT-4o-mini	58.27	**36.79**	62.66	94.69	68.19	7.59	9.59	8.40	**83.42**	34.31	62.61
GPT-4o	47.90	26.15	48.86	93.51	66.65	6.59	9.08	8.10	81.86	28.24	56.17
GPT4.1	43.75	26.68	48.72	92.73	62.47	**6.50**	8.77	**7.03**	79.22	22.49	53.72
Qwen2.5-3b-Instruct-SFT	56.95	20.82	63.30	94.78	65.53	7.74	9.68	8.73	79.12	28.68	58.45
Qwen3-4b-SFT	56.84	31.57	**65.67**	94.69	66.8	8.02	9.53	7.98	82.43	33.09	61.58
+ Refinement: iter=1	**59.12**	28.69	63.50	**94.94**	68.89	7.78	9.26	8.07	82.45	38.01	62.23
+ Refinement: iter=3	56.07	28.53	64.87	94.43	**72.73**	7.84	9.26	9.21	82.68	**43.40**	**63.24**
+ Refinement: iter=5	54.96	28.16	62.70	94.36	69.89	6.57	**8.75**	8.14	83.08	38.57	61.67

Table 2: Evaluation results of our experiments based on 100 validation samples. Refinement here refers to our refinement framework using GPT-4o-mini as our feedback and refinement model. Readability is excluded in the calculation of the average scores.

Model	Relevance					Readability			Clinical		Average
	ROUGE	BLEU	METEOR	BERTScore	Semantic	FKGL	DCRS	CLI	F1CheXbert	F1RadGraph	
Qwen3-4B-SFT	**52.93**	**28.66**	**57.73**	**93.49**	**84.26**	7.53	9.29	8.25	82.69	26.54	**59.01**
+ Refinement: iter=1	52.29	27.84	57.50	93.34	83.70	8.47	9.65	9.09	81.51	**26.84**	58.55
GPT-4o-mini	52.66	26.61	53.92	93.42	82.50	**6.89**	**9.28**	**7.52**	**83.47**	25.83	58.24

Table 3: Evaluation results of selected models across relevance, readability, clinical accuracy, and their averaged metrics based on the test set. Readability is excluded in the calculation of the average scores.

4 Results and Discussion

4.1 Metrics

We use the official evaluation script provided by the organisers (Xiao et al., 2025) to evaluate our models on three aspects: relevance, readability and clinical. Relevance metrics include averaged ROUGE-1, -2, and -L (Lin, 2004) scores, BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2020) and semantic scoring based on SentenceTransformer's fine-tuned MiniLM[1] (Wang et al., 2020). 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) were used to evaluate readability, and F1CheXbert (Smit et al., 2020) and F1RadGraph (Jain et al., 2021) were used for the clinical metrics.

4.2 Results

We used GPT-4o-mini, GPT-4o (OpenAI et al., 2024a) and GPT-4.1[2] as baselines. Due to OpenAI costs, we randomly sampled 100 samples from the validation split to be used for evaluation. To make the results comparable, we performed all our experiments on the 100 samples set. We detail our results for each metric in Table 2.

We calculated the averages of the metrics (excluding the readability metrics) after evaluation to be able to calculate the average of all metrics for each model. From this, we determined that the best performing model was the 3-iteration refinement framework. However, due to limited resources, we submitted the 1-iteration refinement framework instead.

We submitted our fine-tuned Qwen3-4b model and 1-iteration refinement framework for the shared task, along with GPT-4o-mini for our baseline (Table 3). Upon our submissions, we found that the refinement framework underperformed on the test set compared to the fine-tuned model. Calculating the averages of these scores (without the readability metrics) showed that the best model was Qwen3-4b-SFT, which we used as our final submission.

4.3 Analysis

The results in Table 2 show that the refinement framework, particularly when iterations $n = 1$ or $n = 3$, succeeded in improving generations from the fine-tuned model. However, a drop was observed on the test set (Table 3). This section aims to explore possible reasons as to why this had occurred.

4.3.1 Readability and Clinical Metrics

Both tables 2 and 3 show that there is a correlation between the readability metrics and the F1RadGraph metric. To analyse this further, we calculate the correlation between each readability

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[2] https://openai.com/index/gpt-4-1/

Metric	Corr
FKGL vs. F1RadGraph	-0.46
DCRS vs. F1RadGraph	-0.11
CLI vs. F1RadGraph	-0.68

Table 4: Correlations between each readability metric and the F1RadGraph metric (after normalisation).

metric and the F1RadGraph metric (Table 4). From this, a negative correlation can be observed between the readability metrics and F1RadGraph. It can then be inferred that models that scored higher in the F1RadGraph metric tended to have higher readability scores (i.e. produced less readable lay reports). This can be observed in tables 2 and 3, where all the GPT models tended to have better readability scores at the expense of F1RadGraph, and the refinement framework tended to have better F1RadGraph scores at the expense of readability. This is also evidenced by the test set (Table 3), where refinement had the best F1RadGraph scores and the worst readability scores, whereas GPT-4o-mini had the best readability scores but the worst F1RadGraph scores. Our best model on the test set, Qwen3-4B-SFT, was able to balance both readability and F1RadGraph scores.

4.3.2 Affect of Feedback on Refinement Outputs

Madaan et al. (2023) observed that instances where their framework did not improve the original output were primarily caused by erroneous feedback. Therefore, we analysed particular instances within the validation set where using refinement improved on the original generated lay report and where using refinement resulted in worse output to confirm this.

Specific examples are noted in Appendix D. We noticed the feedback model tended to suggest the use of more technical medical terms despite being explicitly instructed that the aim was the generation of lay (i.e. readable) reports, which could affect readability scores. Furthermore, Table 7 shows that poor suggestions could result in less accurate reports (e.g. 'long-term changes' generated by the SFT model vs. 'ongoing changes' generated by the refinement model to describe the term, 'chronic' due to the feedback describing the former as 'vague').

Refined lay reports that achieved higher scores than the initial lay report were those that were accurate but could be written better according to the generated feedback (Tables 8, 9). This implies that refinement works well as an editor for language, but may need fine-tuning on domain data in order to increase factual accuracy.

4.3.3 Lexical Overlap vs. Semantic Overlap

Table 10 in Appendix D shows an example where the refined version of a lay report scored lower than the initial generated report despite being more factually accurate. The term 'interstitial opacities' in the original radiology report could refer to issues such as inflammation or growths; thus, the use of the phrase 'fluid buildup' could be considered an instrinsic hallucination, and the refined report's use of the phrase 'increased density' more faithful to the original radiology report. As metrics such as F1CheXbert and F1RadGraph uses named entity recognition (NER) to evaluate factual accuracy (Smit et al., 2020, Jain et al., 2021), this could lead to bias towards outputs with more overall n-gram overlap with the reference reports. That the refined lay reports that outperformed the initial generated report were primarily those that simply rephrased the initial generated report without changing its meaning (see Section 4.3.2) also supports this hypothesis.

5 Conclusion

By fine-tuning Qwen models, we show that open-source LLMs such as Qwen are capable of generating lay radiology reports that can be easily understood by patients. Despite the refinement framework's performance on the test set, it showed significant results on the validation set and did not underperform the SFT model by a large margin; hence, it has potential for future work. We also analysed potential causes behind the discrepancy in performance between the validation set and the test set. Both approaches exceeded GPT-4o-mini during evaluation, thus proving to be viable approaches in the lay radiology report generation.

Limitations

Due to limited resources, we were unable to utilise the full validation set (which contained 20K samples), which potentially led to a discrepancy when running our models on the full test set. Future work could expand refinement further by experimenting with fine-tuning GPT models and/or open-source LLMs for feedback and refinement to improve performance and increase potential.

Acknowledgements

This research was supported by (1) No. RS-2022-00155911: Artificial Intelligence Convergence Innovation Human Resources Development(Kyung Hee University), (2) No. RS-2024-00509257: Global AI Frontier Lab), and (3) the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2024- 00438239, 35%).

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

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

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Anindya Bijoy Das, Shibbir Ahmed, and Shahnewaz Karim Sakib. 2025. Hallucinations and key information extraction in medical texts: A comprehensive assessment of open-source large language models. *Preprint*, arXiv:2504.19061.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

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.

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, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre 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. *Preprint*, arXiv:2106.14463.

J Peter Kincaid, Robert P Fishburne Jr, 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.

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.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

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.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin,

Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. Gpt-4 technical report. *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 *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. 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. *Preprint*, arXiv:2004.09167.

Bryan D. Steitz, Robert W. Turer, Chen-Tan Lin, Scott MacDonald, Liz Salmi, Adam Wright, Christoph U. Lehmann, Karen Langford, Samuel A. McDonald, Thomas J. Reese, Paul Sternberg, Qingxia Chen, S. Trent Rosenbloom, and Catherine M. DesRoches. 2023. Perspectives of patients about immediate access to test results through an online patient portal. *JAMA Network Open*, 6(3):e233572–e233572.

Nicholas W Sterling, Felix Brann, Stephanie O Frisch, and Justin D Schrager. 2024. Patient-readable radiology report summaries generated via large language model: Safety and quality. *Journal of Patient Experience*, 11:23743735241259477.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Nina S. Vincoff, Matthew A. Barish, and Gregory Grimaldi. 2022. The patient-friendly radiology report: history, evolution, challenges and opportunities. *Clinical Imaging*, 89:128–135.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*.

Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, 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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Kun Zhao, Chenghao Xiao, Chen Tang, Bohao Yang, Kai Ye, Noura Al Moubayed, Liang Zhan, and Chenghua Lin. 2024. X-ray made simple: Radiology report generation and evaluation with layman's terms. *arXiv preprint arXiv:2406.17911*.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. 2024. A survey of large language models in medicine: Progress, application, and challenge. *Preprint*, arXiv:2311.05112.

A Prompts

A.1 Few-shot Prompting For Generation

We used the following prompt to fine-tune our models.

```
### Radiology Report: {example['radiology_report']}
### Layman Report: {example['layman_report']}
```

Figure 2: Prompt used for SFT.

We used a 3-shot prompt to generate lay reports (Figure 3).

```
### You are translating professional radiology
reports into layman's terms. Do not include any
medical jargon. Write concisely. When rewriting the
radiology reports, follow these examples:

Radiology Report: {example[0]['radiology_report']}
Layman's Report: {example[0]['layman_report']}

Radiology Report: {example[1]['radiology_report']}
Layman's Report: {example[1]['layman_report']}

Radiology Report: {example[2]['radiology_report']}
Layman's Report: {example[2]['layman_report']}

### Radiology Report: {radiology_report}
### Layman's Report:
```

Figure 3: 3-shot prompt used for generation

```
### You are an expert medical language reviewer. You are given a radiology report and the full
output generated by a language model in response to it. Evaluate the quality of the **entire
model output** (not just the lay report section) based on the following 3 criteria.

For each, provide a **concise explanation (1-2 sentences max)** and a **score in the format
x/10**. At the end, provide the total score as the **sum of all three criteria**, formatted as
**n/30**.
1. **Factuality (x/10)**: How factually consistent is the output with the original radiology
report? Highlight factually incorrect or inconsistent phrases and penalize accordingly.
2. **Completeness (x/10)**: Does the output include all important information from the radiology
report? Penalize omissions.
3. **Format (x/10)**: Penalize any commentary or non-report language, such as “Here is your
revised report,” “Translation:”, or any explanation of changes. Full marks only if the output
**only** contains the lay summary, without extra headers or commentary.
4. **Total Score (n/30)**: Sum of the seven individual scores.

Here are some examples of evaluations:
Original Radiology Report: \n {examples[0]['radiology_report']}
Lay Report:\n {examples[0]['lay_report']}
Feedback:\n {examples[0]['feedback']}

Original Radiology Report:\n {examples[1]['radiology_report']}
Lay Report:\n {examples[1]['lay_report']}
Feedback:\n {examples[1]['feedback']}

Original Radiology Report:\n {examples[2]['radiology_report']}
Lay Report:\n {examples[2]['lay_report']}
Feedback:\n {examples[2]['feedback']}

## Original Radiology Report:\n {radiology_report}
## Lay Report:\n {lay_report}
## Feedback:
```

Figure 4: Few-shot feedback prompt for the first iteration.

A.2 Feedback and Refinement Prompts

We detail the feedback generation prompts we used in figures 4 and 5. Figure 4 is our few-shot feedback prompt for the single iteration model and the first iteration of the iterative model, whilst Figure 5 is our feedback prompt with past history for subsequent iterations of the iterative model. Figure 6 shows the prompt that includes all seven aspects.

Our prompts for refinement can be seen in figures 7 and 8. Figure 7 is the prompt we use for the first iteration, and Figure 8 is the prompt we use for subsequent iterations.

```
### You are an expert medical language reviewer. You are given a radiology report and the full
output generated by a language model in response to it. Evaluate the quality of the **entire
model output** (not just the lay report section) based on the following 3 criteria.

For each, provide a **concise explanation (1-2 sentences max)** and a **score in the format
x/10**. At the end, provide the total score as the **sum of all three criteria**, formatted as
**n/30**.
1. **Factuality (x/10)**: How factually consistent is the output with the original radiology
report? Highlight factually incorrect or inconsistent phrases and penalize accordingly.
2. **Completeness (x/10)**: Does the output include all important information from the radiology
report? Penalize omissions.
3. **Format (x/10)**: Penalize any commentary or non-report language, such as "Here is your
revised report," "Translation:", or any explanation of changes. Full marks only if the output
**only** contains the lay summary, without extra headers or commentary.
4. **Total Score (n/30)**: Sum of the seven individual scores.

Here are past edits for your reference:
{past_history}

## Original Radiology Report:
{radiology_report}
## Lay Report:
{lay_report}
## Feedback:
```

Figure 5: Few-shot feedback prompt for the first iteration.

```
### You are an expert medical language reviewer. You are given a radiology report and the full
output generated by a language model in response to it. Evaluate the quality of the **entire
model output** (not just the lay report section) based on the following 7 criteria.

For each, provide a **concise explanation (1-2 sentences max)** and a **score in the format
x/10**. At the end, provide the total score as the **sum of all seven criteria**, formatted as
**n/70**.

1. **Factuality (x/10)**: How factually consistent is the output with the original radiology
report? Highlight factually incorrect or inconsistent phrases and penalize accordingly.
2. **Readability (x/10)**: Is the output easy to understand for a patient with no background in
medicine? Identify medical terms or unclear phrasing and penalize as needed.
3. **Completeness (x/10)**: Does the output include all important information from the radiology
report? Penalize omissions.
4. **Conciseness (x/10)**: Is the output concise and succinct? Penalize unnecessarily verbose
outputs (e.g., Outputs that over-explain, or repetitive outputs).
5. **Writing Style (x/10)**: Is the tone formal, objective, and clinical? Penalize conversational
phrasing, direct address (e.g., "you"), or quoting of the original report.
6. **Structure (x/10)**: Does the output follow a clear paragraph-based structure similar to
clinical reports? Penalize if it uses headings, bullet points, or numbered lists.
7. **Format (x/10)**: Penalize any commentary or non-report language, such as "Here is your
revised report," "Translation:", or any explanation of changes. Full marks only if the output
**only** contains the lay summary, without extra headers or commentary.
8. **Total Score (n/70)**: Sum of the seven individual scores.
```

Figure 6: Few-shot feedback prompt for the first iteration.

```
### You are translating radiology reports into layman's terms. You are given feedback for a lay
report. Use the given feedback to improve and rewrite the lay report.
Do not include any commentary, section titles, or explanation of any changes made. The output
should contain only the lay report, written clearly.

### Original Radiology Report: {radiology_report}
### Model Output: {lay_report}
### Feedback: {feedback}
### Use the feedback to improve the lay report. ### Revised Lay Report:
```

Figure 7: Refinement prompt for the first iteration.

```
### You are translating radiology reports into layman's terms. You are given feedback for a lay
report. Use the given feedback to improve and rewrite the lay report.
Do not include any commentary, section titles, or explanation of any changes made. The output
should contain only the lay report, written clearly.

Here are past feedbacks for your reference:
{past_feedbacks}
### Original Radiology Report: {radiology_report}
### Model Output: {lay_report}
### Feedback: {feedback}
### Use the feedback to improve the lay report. ### Revised Lay Report:
```

Figure 8: Refinement prompt for the first iteration.

Generation	Feedback	Refinement	Few-shot	Past History	Iter	Relevance	Readability	Clinical	Avg
Qwen3-4b-FT	Qwen3-4b-FT	Qwen3-4b-FT	✓	✓	3	78.17	88.77	50.00	72.31
Qwen3-4b-FT	GPT-4o-mini	GPT-4o-mini	First	✓	1	88.80	63.38	50.00	67.39
Qwen3-4b-FT	GPT-4o-mini	Qwen3-4b-FT	✓	✗	3	71.99	57.98	50.00	59.99
Qwen3-4b-FT	–	–	–	–	3	71.99	57.98	50.00	59.99
Qwen3-4b-FT	GPT-4o-mini	GPT-4o-mini	First	✓	5	87.69	40.73	50.00	59.47
Qwen3-4b-FT	GPT-4o-mini	Qwen3-4b-FT	First	✗	3	62.91	59.72	50.00	57.54
Qwen3-4b-FT	GPT-4o-mini	GPT-4o-mini	First	✓	3	87.64	30.87	50.00	56.17
Qwen3-4b-FT	Qwen3-4b-FT	Qwen3-4b-FT	All	✗	3	54.12	57.43	50.00	53.85
Qwen3-4b-FT	Qwen3-4b-FT	Qwen3-4b-FT	All	✓	3	55.10	54.51	50.00	53.20
Qwen3-4b-FT	Qwen3-4b-FT	Qwen3-4b-FT	First	✗	3	62.19	36.53	56.25	51.66
Qwen3-4b-FT	GPT-4o-mini	Qwen3-4b-FT	First	✓	3	66.85	28.02	50.00	48.29
Qwen3-4b-FT	GPT-4o-mini	Qwen3-4b-FT	All	✓	3	55.21	36.33	50.00	47.18
Qwen3-4b-FT	GPT-4o-mini	Qwen3-4b-FT	✓	✓	3	55.02	33.05	50.00	46.02
Qwen3-4b-FT	Qwen3-4b-FT	Qwen3-4b-FT	First	✓	3	49.64	29.33	50.00	42.99
Qwen3-4b-FT	Qwen3-4b-FT	Qwen3-4b-FT	✓	✗	3	12.50	49.59	50.00	37.36
Qwen3-4b-FT	GPT-4o-mini	Qwen3-4b-FT	All	✗	3	47.55	8.15	50.00	35.23

Table 5: Comparison of refinement configurations on one validation sample. Few-shot denotes the use of few-shot feedback prompts (First = only on first iteration; All = in every iteration). Past History indicates whether or not past iterations were included in the prompts. Iter is the number of refinement iterations.

B Setting-level Inspection

We also compared the results of different models for feedback and/or refinement with different refinement settings (Table 5).

C Aspect-level Inspection

Aspect	Relevance	Readability	Clinical	Total Average
Completeness	42.77	67.46	89.28	66.50
Factuality	95.11	37.67	50.00	60.93
Format	62.50	64.53	50.00	59.01
Writing Style	61.65	0.00	100.00	53.88
Conciseness	5.44	100.00	50.00	51.81
Structure	17.95	84.39	50.00	50.78
None	43.25	58.90	50.00	50.72
Readability	37.97	61.78	50.00	49.92

Table 6: Average scores across each aspect for relevance, readability, clinical metrics, and overall total on one validation sample.

We inspect the influence each aspect we initially used on the lay report generation. The scores show the average of the evaluation metrics after applying min-max normalisation (Table 6).

D Case Study

We examined specific samples where the initial generation (SFT model) outperformed the refinement framework and vice versa in tables 7, 8, 9 and 10 for the analyses in sections 4.3.2 and 4.3.3.

Original Radiology Report	First Generation	Feedback	Refined
Cardiomegaly. Aortic elongation. Chronic parenchymal lung changes.	An enlarged heart is seen. The main blood vessel (aorta) is longer than usual. There are long-term changes in the lung tissue.	1. **Factuality (7/10)**: The output accurately reflects the findings of the original report, but the phrase "long-term changes in the lung tissue" could be interpreted as vague compared to "chronic parenchymal lung changes." This slight discrepancy in terminology affects the factuality score. 2. **Completeness (8/10)**: The lay report captures all key findings from the original report, but it could be more specific regarding the nature of the lung changes. The term "parenchymal" is omitted, which is a significant descriptor in the context of lung changes. 3. **Format (9/10)**: The output is mostly well-formatted, containing only the lay summary without additional commentary or headers. However, the phrase "is seen" in the first sentence could be considered unnecessary and slightly detracts from the directness of the report. 4. **Total Score (24/30)**: The total score reflects the evaluation across all three criteria, indicating a generally good quality of the model output with some areas for improvement.	The heart appears larger than normal. The major blood vessel leaving the heart, called the aorta, is stretched out. There are ongoing changes in the breathing part of the lungs.

Table 7: Sample where refinement performed worse in the evaluation metrics than the first generation due to feedback.

Original Radiology Report	First Generation	Feedback	Refined
Cardiomegaly. Pulmonary parenchyma without significant alterations. Correlate clinically.	The heart appears larger than normal. The lung tissue looks fine without major issues. Discuss this with your doctor for further evaluation.	1. **Factuality (8/10)**: The output accurately conveys the essence of the original report, but the phrase "the heart appears larger than normal" could be more precise by using the term "cardiomegaly" directly. Additionally, "lung tissue looks fine" lacks the nuance of "without significant alterations," which could imply a more technical assessment. 2. **Completeness (7/10)**: The lay report captures the main findings but omits the recommendation to "correlate clinically," which is an important aspect of the original report indicating the need for further clinical correlation. 3. **Format (9/10)**: The output is presented in a clear and straightforward manner without unnecessary commentary or headers, but the phrase "Discuss this with your doctor for further evaluation" could be seen as slightly extraneous, as it doesn't directly reflect the original report's wording. 4. **Total Score (24/30)**: The overall output is mostly accurate and well-formatted, with minor issues in factual precision and completeness.	The heart is enlarged. The lungs look normal without any major problems. Work with your doctor to discuss these results.

Table 8: Sample where refinement performed better in the evaluation metrics than the first generation.

Original Radiology Report	First Generation	Feedback	Refined
No pulmonary alterations suggestive of lung infection or other relevant abnormalities are observed.	The lungs look fine, showing no signs of infections or other serious problems.	1. **Factuality (8/10)**: The lay report accurately reflects the findings of the original radiology report, stating that there are no signs of infection or serious problems in the lungs. However, the phrase "the lungs look fine" could be seen as overly simplistic and not as precise as the original wording, which may warrant a slight deduction. 2. **Completeness (10/10)**: The lay report includes all important information from the original report, effectively conveying that there are no pulmonary alterations or relevant abnormalities. 3. **Format (8/10)**: While the lay report is mostly straightforward, it includes the phrase "The lungs look fine," which could be interpreted as informal commentary. A more direct paraphrase of the original report would have been preferable for a formal report format. 4. **Total Score (26/30)**: The overall score reflects a strong performance, with minor deductions for factual precision and format.	There are no signs of lung infections or other significant issues in the lungs.

Table 9: Sample where refinement performed better in the evaluation metrics than the first generation.

Original Radiology Report	First Generation	Refined	Reference
Findings in the lung parenchyma show interstitial opacities with a diffuse distribution throughout the lung fields, particularly in the middle and lower regions of the right lung and the middle and lower regions of the left lung. Conclusion: Findings are consistent with COVID-19.	The lungs show signs of fluid buildup between the air sacs, spread all over both sides of the lungs, especially in the middle and bottom parts of the right lung and the middle and bottom parts of the left lung. Conclusion: These findings match what we see in people with COVID-19.	The images show areas of increased density within the lung tissue that are evenly distributed across both lungs, mainly affecting the central and lower sections on both sides. It looks like these patterns are typical of what you would expect to see in someone with COVID-19.	The examination of the lung tissue shows a widespread pattern of haziness affecting the lung fields, especially in the middle and lower parts of both the right and left lungs. Conclusion: These findings are typical of what is seen in COVID-19.

Table 10: Sample where refinement performed worse in the evaluation metrics than the first generation despite being more factually accurate. The highlighted phrases represent different interpretations of the same finding from the original radiology report with varying degrees of faithfulness to the original text, as mentioned in Section 4.3.3.