Enhancing Radiology Report Summarizations Through Performance Feedback

Varun Venkat Rao
University of Michigan - EECS Department
Ann Arbor, MI
varu@umich.edu

Abstract

Large Language Models (LLMs), such as those in the GPT and BART families, have demonstrated exceptional prowess in capturing and condensing crucial contextual information, achieving state-of-the-art performance in summarization tasks. However, growing concerns within the community revolve around the hallucination issues exhibited by these models. In particular, LLMs can sometimes generate factually incorrect summaries, a concern of utmost significance in clinical domain Natural Language Processing (NLP) tasks, like radiology report summarization, where inaccuracies can lead to severe diagnostic errors. Fine-tuning LLMs using reinforcement learning techniques, such as learning from human feedback and direct preference optimization, has shown promise in aligning these models to generate factually consistent summaries. However, such training procedures traditionally necessitate large quantities of human-annotated data, posing a significant cost challenge. In this work, we propose a novel pipeline that leverages fine-tuned language models instead of human experts to generate feedback data. This approach aims to enhance the quality of generated radiology summaries. Furthermore, we demonstrate the effectiveness of incorporating feedback examples generated through this pipeline when aligning LLMs with the task of radiology report summarization. Our evaluation encompasses performance assessments against various baselines on both in-domain and out-of-domain datasets. We highlight that our proposed approach yields notable improvements across all tested metrics, boasting a 4.97%, 3.05%, and 23.83% enhancement in RougeL, F1-RadGraph, and F1-Chexbert metrics, respectively.

1 Introduction

In the realm of healthcare, where precision and efficiency are paramount, the documentation of medical findings stands as a critical component. Radiology reports, in particular, demand concise yet comprehensive summarization to distill essential insights from extensive textual information. Our focus centers on the 'Impression sections' derived from the Findings and/or Background portions of radiology reports, aiming to streamline the arduous task clinicians face in summarizing intricate details from radiological investigations [Fleming u. a. (2023); Arndt u. a. (2017); Van Veen u. a. (2023); Golob Jr u. a. (2016)].

The current state of radiology report summarization poses challenges, especially in terms of time consumption and potential errors introduced during the manual summarization process. Even highly experienced physicians find themselves grappling with the intricacies of this task, emphasizing the need for innovative solutions in a field where accuracy is non-negotiable [Bowman (2013); Johnson u. a. (2020); Yackel und Embi (2010); Van Veen u. a. (2023)].

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Compounding this issue, low-resource communities face a distinct disadvantage due to limited access to expert radiologists, exacerbating the challenge of producing accurate radiology reports. In response to this, our research addresses the pressing need for a reliable tool capable of generating precise and succinct summaries, not only to alleviate the workload on expert radiologists but also to contribute to the overall well-being of healthcare professionals and enhance the quality of healthcare delivery.

Recent years have witnessed the rise of Large Language Models (LLMs) such as T5, GPT, BART, and LLaMA-2, showcasing significant advancements in natural language understanding and generation tasks. However, a persistent challenge lies in the tendency of these models to generate hallucinations, resulting in factually inconsistent outputs [Ji u. a. (2023); OpenAI (2023); Zhang u. a. (2023); Maynez u. a. (2020)]. Aligning these LLMs to ensure factual consistency has become a focal point of research, particularly in generation tasks [Shi u. a. (2023); Cao u. a. (2021); Kang und Hashimoto (2020); Goyal u. a. (2023)].

This paper contributes to this evolving landscape by presenting a novel approach to radiology report summarization. We delve into the fine-tuning of pre-trained LLMs such as GPT2, BART, and Llama, comparing their performance against clinical counterparts (BioGPT, BioBART) specifically pre-trained on vast medical text corpora. Notably, our work extends beyond model performance evaluation; we introduce a preference dataset for radiology report summarization based on performance metrics for reinforcement learning alignment tasks. This innovation eliminates the dependence on expert human annotators, thereby addressing the scarcity and expense associated with obtaining expert-level annotations in clinical domains.

Furthermore, we leverage Direct Preference Optimization (DPO), a model alignment technique, to enhance the quality of summaries generated by the fine-tuned models. Our approach aims to bridge the gap between advanced language models and the critical need for accurate and reliable radiology report summarization, providing a promising solution to the challenges faced in healthcare documentation. In the subsequent sections, we delve into the methodology, results, and implications of our research, highlighting its potential to significantly impact both clinical practice and the broader landscape of natural language processing in healthcare.

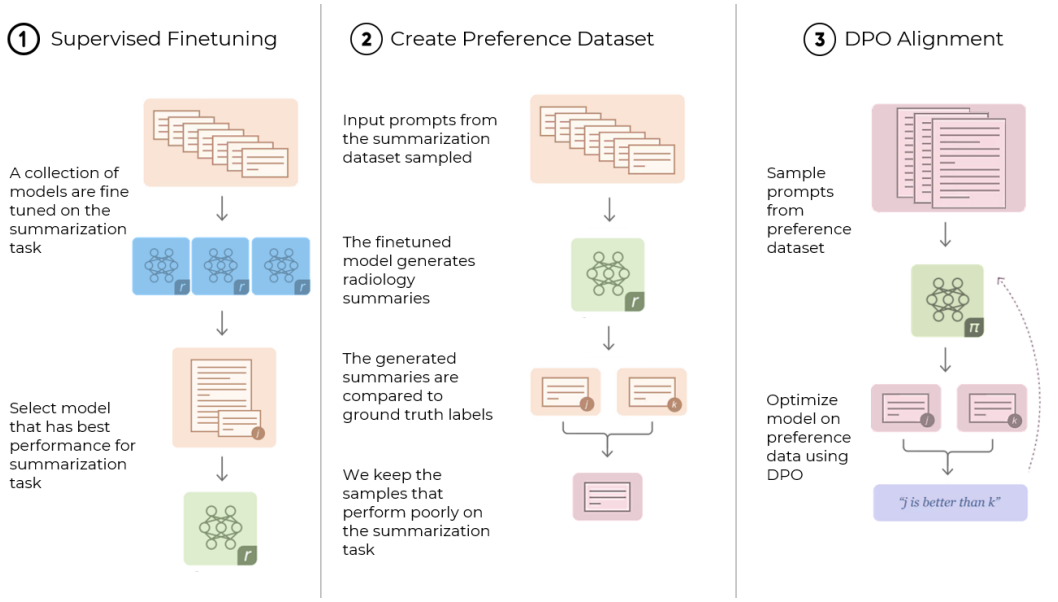

Figure 1: High-level workflow: (1) Tailor transformer-based models for radiology report summarization through supervised learning. (2) Construct a preference dataset using the fine-tuned model, showcasing preferred and undesired summaries. (3) Refine model alignment with preferred summaries using Direct Preference Optimization (DPO) in a reinforcement learning framework

2 Related Work

2.1 Transformer-Based Language Models in NLP

In recent years, transformer-based language models (LMs) have revolutionized natural language processing (NLP) with their state-of-the-art performance across various tasks such as language generation, question answering, and machine translation [Vaswani u. a. (2017)]. Pioneering models like BERT [Devlin u. a. (2018)] and GPT-2 [Radford u. a. (2019)] introduced a paradigm shift by first training on large amounts of general data and subsequently fine-tuning on domain-specific data. This approach has paved the way for the development of increasingly sophisticated large language models (LLMs) such as GPT-3 [Brown u. a. (2020)], PaLM [Chowdhery u. a. (2023)], and T5 [Raffel u. a. (2020)].

However, when applied to the clinical domain, these LMs and LLMs often grapple with a lack of medical knowledge, leading to a proliferation of factual errors [Petroni u. a. (2019); Sung u. a. (2021); Yao u. a. (2022)]. To address this, a common strategy involves continuing the pretraining of LMs on domain-specific data, typically derived from PubMed articles and clinical notes. Notably, Luo u. a. (2022) proposed an alternative approach, advocating for pretraining on biomedical domain-specific data from scratch, highlighting the importance of a vocabulary more aligned with the target biomedical domain. Similarly, Wu u. a. (2023) introduced the LLama 2 model, pre-trained on 4.8M biomedical academic papers and 30K medical books, demonstrating improved performance on various biomedical literature language processing tasks compared to their general domain counterparts.

2.2 Learning with Feedback Paradigms

While domain-specific language models outperform their general domain counterparts, challenges persist in the standard Supervised Fine-Tuning (SFT) approach. Stiennon u. a. (2020) emphasize that SFT treats important errors and unimportant errors alike in the loss function, limiting the model's ability to consistently produce high-quality, human-determined text, particularly in terms of factuality. Recognizing this, recent work has explored learning paradigms with human feedback [Böhm u. a. (2019); Ziegler u. a. (2019); Stiennon u. a. (2020); Akyürek u. a. (2023); Dong u. a. (2023);Zhao u. a. (2023); Yuan u. a. (2023)].

Reinforcement Learning from Human Feedback (RLHF) has gained prominence as a technique for aligning large language models with human preferences. This approach involves training models within a reinforcement learning framework, utilizing reward signals derived from human-generated feedback on model outputs. Despite its promise, RLHF faces challenges due to the substantial need for human-annotated preference data, especially in domains where such data is scarce or expensive.

Direct Preference Optimization (DPO) emerges as a compelling alternative to RLHF, addressing the limitations associated with heavy reliance on annotated preference data. DPO optimizes models based on direct comparisons of multiple model-generated outputs, offering a pragmatic solution to the challenges posed by RLHF. This approach proves particularly advantageous in tasks like text summarization, where obtaining human preferences is often more feasible than detailed annotated data.

DPO's comparative advantage lies in its potential to improve model alignment and generate outputs that are not only factually consistent but also closely aligned with human preferences. This proves especially crucial in mitigating challenges associated with hallucinations and inconsistent outputs from large language models, as identified in recent advancements [Ji u. a. (2023); OpenAI (2023); Zhang u. a. (2023); Maynez u. a. (2020)]. As we delve into the application of DPO in the subsequent sections, our work aims to contribute to the ongoing discourse on refining text summarization models for improved reliability and utility in real-world applications.

3 Dataset

The primary dataset employed in this study is the MIMIC-CXR dataset, a rich collection comprising 125,417 multimodal triplets. Each triplet includes chest X-ray images, corresponding findings, and the impression section extracted from radiology reports. The fundamental objective of our research is to address the Radiology Report Summarization (RRS) task, specifically focusing on generating

Table 1: Number of reports in the MIMIC-CXR dataset and hidden CheXpert Dataset.

Modality/Anatomy	Number of Images	Number of reports	Train	Validation	Test
MIMIC Chest X-Rays	237564	128032	125417	991	1624
CheXpert Chest X-Rays (hidden)	1000	1000	-	-	1000

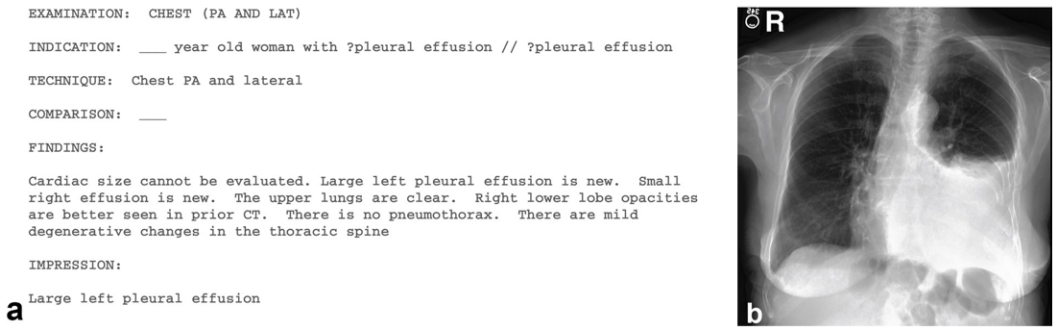

Figure 2: Example study contained in MIMIC-CXR. Left (a), the radiology report provides the interpretation of the image. PHI has been removed and replaced with three underscores ___. Right, the chest radiographs for this study are shown: (b) the frontal view [Johnson u. a. (2019)]

concise and comprehensive summaries (Impression/labels) from the Findings and/or Background sections of radiology reports.

The MIMIC-CXR dataset is sourced from two reputable repositories: PhysioNet [Johnson u. a. (2020)] and ViLMedic [Delbrouck u. a. (2022)]. These repositories provided access to pre-processed versions of the MIMIC-CXR dataset, ensuring the removal of any confidential patient information and adhering to ethical standards in the handling of medical data.

Additionally, our research incorporates the use of a new out-of-domain test set from Stanford's CheXpert [Irvin u. a. (2019)]. This particular set from Stanford forms part of the hidden test set specifically tailored to measure our model's robustness and generalization performance.

4 Approach

4.1 High-level methodology

Our methodology closely follows the framework outlined in Rafailov u. a. (2023), with adaptations tailored to our synthetic preference dataset. The entire process, depicted in Figure 1, unfolds in three key steps.

Step 1: Fine-tune models through supervised learning on a designated dataset. In this initial phase, we carefully choose a diverse array of transformer-based Large Language Models (LLMs). Our focus is on refining these models through supervised learning, specifically tailoring them for the task of radiology report summarization, with the overarching goal of generating concise and factually accurate impressions from the findings section of a radiology report.

Step 2: Construct a preference dataset by employing the previously fine-tuned model. Given an LM finetuned on the summarization task, our next step involves crafting a preference dataset. This dataset comprises examples showcasing both a preferred output summary and an undesired output summary for a given report.

Step 3: Utilize Direct Preference Optimization (DPO) to refine the model alignment with preferred summaries. We use the output pairs from our newly constructed dataset as reward signals used to optimize a reinforcement learning algorithm, specifically the DPO algorithm.

4.2 Supervised finetuning

We investigate a diverse collection of transformer-based LLMs for clinical summarization tasks. This includes two broad approaches to language generation: sequence-to-sequence (seq2seq) models and autoregressive models. Seq2seq models use an encoder-decoder architecture to map the input text to a generated output, often requiring paired datasets for training. These models have shown strong performance in machine translation [Chen u. a. (2018)] and summarization [Shi u. a. (2021)]. In contrast, autoregressive models' architecture uses only a decoder. They generate tokens sequentially—where each new token is conditioned on previous tokens—thus efficiently capturing context and long-range dependencies. Autoregressive models are typically trained with unpaired data, and they are particularly useful for NLP tasks such as text generation, question-answering, and dialogue interactions [Brown u. a. (2020); Chiang u. a. (2023)]. We include prominent seq2seq models due to their strong summarization performance [Shi u. a. (2021)] and autoregressive models due to their state-of-the-art performance for many general NLP tasks [Zheng u. a. (2023)]. As shown in Table 2, our choice of models can be organized into four main categories: (1) General domain seq2seq models (2) General domain autoregressive models (3) Clinical domain seq2seq models, and (4) Clinical domain autoregressive models.

Table 2: Quantitative Selection of Seven Models for Radiology Report Summarization, Including State-of-the-Art Sequence-to-Sequence and Autoregressive Models.

Model	Num. of Parameters	Seq2Seq	Autoreg.	General Domain	Clinical Domain
BART$_{base}$	140M	✔	-	✔	-
BART$_{large}$	400M	✔	-	✔	-
BioBART$_{base}$	140M	✔	-	-	✔
BioBART$_{large}$	400M	✔	-	-	✔
GPT-2$_{medium}$	355M	-	✔	✔	-
GPT-2$_{large}$	812M	-	✔	✔	-
BioGPT$_{base}$	347M	-	✔	-	✔

This step serves a dual purpose: firstly, it discerns the optimal models for effective execution in clinical summarization tasks, and secondly, it incorporates a supervised fine-tuning process that facilitates the model's grasp of fundamental task intricacies. This preparatory phase is indispensable before engaging in reinforcement learning. By instilling a foundational level of understanding, the model is endowed with the capability to formulate informed predictions prior to exposure to reinforcement signals. This strategic initialization mitigates the exploration challenge within the reinforcement learning framework. With some prior knowledge in place, the model is better equipped to navigate potential actions, thereby enhancing the efficiency and effectiveness of the exploration process. These considerations collectively contribute to the refinement of the reinforcement learning framework, ensuring the model's readiness for subsequent phases of adaptation and optimization in clinical summarization tasks.

4.3 Preference-data generation

In the context of summarization alignment through Direct Preference Optimization (DPO), the model derives insights from a pivotal preference dataset, denoted as $D_{pref} : \{X, Y_+, Y_-\}$, where Y_+ represents a collection of preferred summaries, and Y_- comprises the less favored ones. The model fine-tunes its learning by seeking to elevate the probability of generating summaries aligned with Y_+ while concurrently reducing the likelihood of generating summaries resembling Y_-. The process of generating Y_+ proves relatively straightforward, as it involves extracting insights directly from our ground truth labels embedded in the reference summaries within our original dataset D. In contrast, generating samples for Y_- often necessitates the participation of human annotators, rendering it a more resource-intensive and challenging undertaking.

To construct dispreferred summaries, indicated as y_-^i within the collection Y_-, we capitalize on the capabilities of our fine-tuned model from the prior stage. This sophisticated model generates

summaries, y_i, by processing clinical reports extracted from our initial dataset, D. Following this generation phase, we subject these summaries to a thorough quality assessment, employing widely recognized summarization evaluation metrics such as RougeL, Blue, Bertscore, F1Radgraph, and F1CheXbert.

In essence, we undertake a meticulous filtering process based on the model's performance. Summaries showcasing exceptional proficiency, as evidenced by favorable metrics, are deliberately excluded. Our attention shifts to retaining clinical reports where the model manifests suboptimal summarization outcomes. More precisely, we focus on report instances where the model's generated summaries scored below 30% in both F1RadGraph and RougeL metrics. These specific instances serve as the cornerstone for constructing our dispreferred summaries, Y_-, a pivotal element within our Direct Preference Optimization (DPO) preference dataset, D_{pref}. This strategic methodology guarantees the development of a robust dataset, instrumental in the subsequent optimization of our model.

4.4 Preference-based training

The last step of our approach involves using this synthetically generated preference data D_{pref} : $\{X, Y_+, Y_-\}$ to fine-tune a π_{ref} using DPO to generate factually consistent outputs.

$$l_{dpo}(\pi_\theta; \pi_{ref}) = -E_{(x^i, y_+^i, y_-^i) \sim D_{pref}} \left(\log \sigma \left[\beta \log \frac{\pi_\theta(y_+^i|x^i)}{\pi_{ref}(y_+^i|x^i)} - \beta \log \frac{\pi_\theta(y_-^i|x^i)}{\pi_{ref}(y_-^i|x^i)} \right] \right) \quad (1)$$

For aligning π_{ref} using DPO ($\pi_{ref} \rightarrow \pi_\theta$) we train the model by optimizing the loss function l_{dpo} shown in equation 1, where given the preference data D_{pref} : $\{X, Y_+, Y_-\}$ consisting of a set of clinical notes x^i, preferred summaries Y_+^i (ground truth reference summaries), and the dispreferred summaries Y_-^i (hallucinated summaries), the mode learns to increase the likelihood of the preferred summaries Y_+^i and to decrease the likelihood of the dispreferred summaries Y_-^i. In the equation, π_{ref} is the base model and π_θ is the model being trained to have improved alignment and β is used to scale the weight on how incorrect the model should treat the dispreferred summary y_-^i relative to the preferred summary y_+^i. The higher the β beta, the less the divergence from the initial policy π_{ref}.

4.5 Evaluation Method

To assess the efficiency of our supervised finetuned (SFT) models and our reinforcement learning aligned models, we employ a battery of evaluation metrics, both for the MIMIC-CXR and Chexpert datasets, including ROUGE-L, F1-RadGraph, and F1-CheXbert. Now while the BLEU metric is a pretty well-known simple summarization metric, we chose to use the ROUGE-L metric that evaluates similarity based on the longest common subsequence; it considers both precision and recall, hence being more comprehensive than BLEU.

F1-RadGraph and F1-CheXbert are both F-score style metrics that measure the factual correctness, consistency, and completeness of generated radiology reports compared to the reference. F1-RadGraph uses RadGraph [Jain u. a. (2021)], a graph dataset of entities and relations present in radiology reports, to calculate the overlap in the extracted clinical entities and relations. F1-CheXbert [Delbrouck u. a. (2022)] on the other hand uses a rule-based labeler that searches radiological reports for mentions of the conditions and predicts a label from 14 diagnostic classes derived from the CheXpert dataset (12 unique conditions/classes, along with an uncertainty and blank label).

5 Experiments and Results

5.1 Supervised finetuning

We fine-tune each model listed in Table 2 using data from the training split of the MIMIC-CXR dataset. The training process for each model is executed on a single NVIDIA A40 GPU, and we adhere to the following set of hyperparameters during the process:

- Initial learning rate of $2e^{-5}$ that linearly decays after a 100-step warm-up

- Five epochs maximum with an early stopping criterion if the validation loss has not decayed for 2 consecutive epochs
- Batch size of 16 for large architectures (like GPT-2_{large}) or 32 for base architectures with 16 gradient accumulation steps, rendering an effective batch size of 256 or 512 respectively.

To assess the effectiveness of our supervised models, we employ the evaluation metrics detailed in Section 4.5. These metrics are then used to compute scores for each of our models on the test split of the MIMIC-CXR dataset. The obtained results are presented in Table 3 for reference.

Table 3: Comparing Supervised Finetuned Models on the MIMIC-CXR Test Split: Assessing Performance Differences Between Seq2Seq and Autoregressive Architectures in Both General and Clinical Domains

Architecture	Model	Parameters	RougeL	F1-RadGraph	F1-CheXbert
Seq2Seq	BART$_{base}$	140M	41.45	38.58	57.59
	BART$_{large}$	400M	42.7	39.28	58.43
	BioBART$_{base}$	140M	40.68	37.5	57.79
	BioBART$_{large}$	400M	**43.13**	<u>39.511</u>	<u>59.16</u>
Autoreg.	GPT2$_{medium}$	355M	35.38	33.97	60.49
	GPT2$_{large}$	812M	40.17	38.15	68.45
	BioGPT$_{base}$	347M	<u>42.93</u>	**40.70**	**71.55**

Examining Table 3 reveals a substantial performance superiority of clinical domain models over their general domain counterparts. Particularly noteworthy are the BioGPT and BioBART$_{large}$ models, standing out as the top performers within their respective architecture types (autoregressive and seq2seq). Furthermore, there is a prevailing trend evident in the results, indicating that larger model variants consistently exhibit superior performance compared to their base counterparts. The sole exception to this trend is observed in the BioGPT and BioBART models, both equipped with around 400M parameters, surpassing the GPT2-large model boasting approximately 812M parameters. This anomaly is very interesting as it underscores the prowess of these clinical domain models.

When evaluating the performances of the BioGPT and BioBART models, it becomes evident that although the RougeL score of the BioGPT model slightly trails that of the BioBART model, this difference is mitigated by the BioGPT model achieving higher scores in both the F1-RadGraph and F1-CheXbert metrics. Notably, despite the BioGPT and BioBART models exhibiting closely aligned RougeL and F1-RadGraph scores, the former demonstrates a substantial 12.39% enhancement in the F1-CheXbert metric. This significant improvement accentuates the BioGPT model's efficacy in capturing nuanced aspects, thereby contributing to its noteworthy overall performance.

5.2 RL alignment with metrics preference

In the reinforcement learning phase, our emphasis is on aligning the previously fine-tuned models discussed in Section 5 with the preference dataset established through the procedures outlined in Section 4.3 (an illustration of a sample from this dataset is presented in Figure 3). Specifically, we will employ the BioGPT and BioBART$_{large}$ models as our baseline models for this experiment. The alignment procedure for each model was conducted on a single NVIDIA A40 GPU, employing the following set of hyperparameters:

- Initial learning rate of $5e^{-}7$ without any decays
- Trained for a single epoch on the DPO preference dataset
- Batch size of 16 was used for both models with 16 gradient accumulation steps, rendering an effective batch size of 256
- DPO beta was set to 0.1
- Maximum prompt length of 512 tokens and Max sequence length of 1024 tokens

Prompt	The heart is normal in size. The mediastinal and hilar contours appear within normal limits. Each hilum is mildly prominent, probably suggesting mild prominence of central pulmonary vessels, but there is no frank congestive heart failure. No focal opacification is seen aside from streaky left lower lung opacity suggesting minor atelectasis. There is no pleural effusion or pneumothorax. Bony structures are unremarkable. The main impression based on the given FINDINGS section of the chest X-ray report are:
Chosen	Mild perihilar prominence, suspected to represent mildly prominent pulmonary vessels without definite pneumonia. Streaky left basilar opacification seen only on the frontal view is probably due to minor atelectasis or scarring.
Rejected	No evidence of acute disease.

Figure 3: Illustration from the synthesized preference dataset utilized for DPO alignment. (Top) Depicts the input prompt provided to the models, (Middle) Showcases an instance of the preferred output summary corresponding to the provided prompt, (Bottom) Exhibits an example of a dispreferred or rejected output summary.

After the alignment phase, we assessed the performance of the SF baseline models and DPO-aligned models on the test splits of both the MIMIC-CXR and CheXpert datasets to evaluate the effectiveness of our alignment approach. The results are presented in Table 4 and Table 5 below.

Table 4: Comparative Assessment: DPO Finetuned Models Versus SFT Models in MIMIC-CXR Test Split Evaluation

Dataset	Model	DPO	RougeL	F1-RadGraph	F1-CheXbert
MIMIC-CXR	BioBART$_{large}$	-	43.13	39.511	59.16
	BioBART$_{large}$	✔	**45.87**	**42.56**	**73.59**
	BioGPT$_{base}$	-	42.9328	40.695	71.549
	BioGPT$_{base}$	✔	42.8901	<u>40.7084</u>	71.5036

Table 5: Evaluating Out-of-Domain (OOD) Model Performance: A Comparison Between DPO Models and SFT Models on the OOD Test Dataset (Chexpert)

Dataset	Model	DPO	RougeL	F1-RadGraph	F1-CheXbert
CheXpert	BioBART$_{large}$	-	26.28	10.56	39.49
	BioBART$_{large}$	✔	<u>31.25</u>	<u>13.34</u>	<u>63.32</u>
	BioGPT$_{base}$	-	29.90	13.10	63.34
	BioGPT$_{base}$	✔	**32.75**	**14.02**	**68.83**

Upon comparing the outcomes of the DPO and SFT models on the MIMIC-CXR test set, the performance of the BioGPT-DPO model remains relatively stable, with a slight reduction in RougeL and F1-CheXbert scores. In contrast, the BioBART model, aligned using DPO, exhibits substantial enhancements in model performance across all evaluated metrics. Specifically, there is a noteworthy 2.74% surge in RougeL scores, a commendable 3.05% improvement in F1-Radgraph scores, and an impressive 14.43% boost in F1-CheXbert scores compared to its SFT baseline counterparts.

Moreover, when assessing these models on out-of-domain test data (CheXpert), both DPO models outshine their SFT counterparts. It's intriguing to note that, despite the BioGPT-DPO model exhibiting

diminished performance on the in-domain test split, it emerges as the top-performing model when evaluated on this out-of-domain dataset. Lastly, we emphasize that although the BioBART-DPO model showed slightly inferior performance compared to the BioGPT-DPO model, it demonstrated significant improvement compared to the BioBART-SFT model. Notably, we observe a 4.97% enhancement in RougeL performance, a 2.78% improvement in F1-Radgraph scores, and a remarkable 23.83% advancement in F1-CheXbert scores.

This notable performance surge is especially remarkable given that the DPO models were fine-tuned with just around 10,000 samples for a single epoch. In stark contrast, the SFT models underwent extensive training with over 120,000+ samples for over 5 epochs, yet displayed only marginal improvements at each step, approximately less than 0.5%. Our proposed models demonstrate a substantial enhancement in model performance, both on in-domain and out-of-domain test set data, all while utilizing only 1.6% of the data required by SFT models to achieve comparable yet inferior performance.

6 Conclusion

In this paper, we conduct a comprehensive analysis of various autoregressive and seq2seq models for radiology report summarization, showcasing the superior capabilities of clinical domain models when compared to their general domain counterparts. The main focus of our study however is to illustrate the effectiveness of leveraging performance-based feedback examples to enhance factual alignment in Language Models (LLMs) dedicated to radiology report summarization. Additionally, we also introduce a novel pipeline for generating synthetic preference-based data, eliminating the reliance on human annotators. Through our proposed approach, we observe outstanding performance on both in-domain and out-of-domain test sets. This includes a notable 4.97% improvement in RougeL scores, a substantial 3.05% enhancement in F1-Radgraph scores, and an impressive 23.83% boost in F1-CheXbert scores.

References

[Akyürek u. a. 2023] AKYÜREK, Afra F. ; AKYÜREK, Ekin ; MADAAN, Aman ; KALYAN, Ashwin ; CLARK, Peter ; WIJAYA, Derry ; TANDON, Niket: RL4F: Generating Natural Language Feedback with Reinforcement Learning for Repairing Model Outputs. In: *arXiv preprint arXiv:2305.08844* (2023)

[Arndt u. a. 2017] ARNDT, Brian G. ; BEASLEY, John W. ; WATKINSON, Michelle D. ; TEMTE, Jonathan L. ; TUAN, Wen-Jan ; SINSKY, Christine A. ; GILCHRIST, Valerie J.: Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. In: *The Annals of Family Medicine* 15 (2017), Nr. 5, S. 419–426

[Böhm u. a. 2019] BÖHM, Florian ; GAO, Yang ; MEYER, Christian M. ; SHAPIRA, Ori ; DAGAN, Ido ; GUREVYCH, Iryna: Better rewards yield better summaries: Learning to summarise without references. In: *arXiv preprint arXiv:1909.01214* (2019)

[Bowman 2013] BOWMAN, Sue: Impact of electronic health record systems on information integrity: quality and safety implications. In: *Perspectives in health information management* 10 (2013), Nr. Fall

[Brown u. a. 2020] BROWN, Tom ; MANN, Benjamin ; RYDER, Nick ; SUBBIAH, Melanie ; KAPLAN, Jared D. ; DHARIWAL, Prafulla ; NEELAKANTAN, Arvind ; SHYAM, Pranav ; SASTRY, Girish ; ASKELL, Amanda u. a.: Language models are few-shot learners. In: *Advances in neural information processing systems* 33 (2020), S. 1877–1901

[Cao u. a. 2021] CAO, Meng ; DONG, Yue ; CHEUNG, Jackie Chi K.: Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In: *arXiv preprint arXiv:2109.09784* (2021)

[Chen u. a. 2018] CHEN, Mia X. ; FIRAT, Orhan ; BAPNA, Ankur ; JOHNSON, Melvin ; MACHEREY, Wolfgang ; FOSTER, George ; JONES, Llion ; PARMAR, Niki ; SCHUSTER, Mike ; CHEN, Zhifeng u. a.: The best of both worlds: Combining recent advances in neural machine translation. In: *arXiv preprint arXiv:1804.09849* (2018)

[Chiang u. a. 2023] CHIANG, Wei-Lin ; LI, Zhuohan ; LIN, Zi ; SHENG, Ying ; WU, Zhanghao ; ZHANG, Hao ; ZHENG, Lianmin ; ZHUANG, Siyuan ; ZHUANG, Yonghao ; GONZALEZ, Joseph E. u. a.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. In: *See https://vicuna. lmsys. org (accessed 14 April 2023)* (2023)

[Chowdhery u. a. 2023] CHOWDHERY, Aakanksha ; NARANG, Sharan ; DEVLIN, Jacob ; BOSMA, Maarten ; MISHRA, Gaurav ; ROBERTS, Adam ; BARHAM, Paul ; CHUNG, Hyung W. ; SUTTON, Charles ; GEHRMANN, Sebastian u. a.: Palm: Scaling language modeling with pathways. In: *Journal of Machine Learning Research* 24 (2023), Nr. 240, S. 1–113

[Delbrouck u. a. 2022] DELBROUCK, Jean-benoit ; SAAB, Khaled ; VARMA, Maya ; EYUBOGLU, Sabri ; CHAMBON, Pierre ; DUNNMON, Jared ; ZAMBRANO, Juan ; CHAUDHARI, Akshay ; LANGLOTZ, Curtis: ViLMedic: a framework for research at the intersection of vision and language in medical AI. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2022, S. 23–34

[Devlin u. a. 2018] DEVLIN, Jacob ; CHANG, Ming-Wei ; LEE, Kenton ; TOUTANOVA, Kristina: Bert: Pre-training of deep bidirectional transformers for language understanding. In: *arXiv preprint arXiv:1810.04805* (2018)

[Dong u. a. 2023] DONG, Hanze ; XIONG, Wei ; GOYAL, Deepanshu ; PAN, Rui ; DIAO, Shizhe ; ZHANG, Jipeng ; SHUM, Kashun ; ZHANG, Tong: Raft: Reward ranked finetuning for generative foundation model alignment. In: *arXiv preprint arXiv:2304.06767* (2023)

[Fleming u. a. 2023] FLEMING, Scott L. ; LOZANO, Alejandro ; HABERKORN, William J. ; JINDAL, Jenelle A. ; REIS, Eduardo P. ; THAPA, Rahul ; BLANKEMEIER, Louis ; GENKINS, Julian Z. ; STEINBERG, Ethan ; NAYAK, Ashwin u. a.: Medalign: A clinician-generated dataset for instruction following with electronic medical records. In: *arXiv preprint arXiv:2308.14089* (2023)

[Golob Jr u. a. 2016] GOLOB JR, Joseph F. ; COMO, John J. ; CLARIDGE, Jeffrey A.: The painful truth: The documentation burden of a trauma surgeon. In: *Journal of Trauma and Acute Care Surgery* 80 (2016), Nr. 5, S. 742–747

[Goyal u. a. 2023] GOYAL, Navita ; NENKOVA, Ani ; DAUMÉ III, Hal: Factual or Contextual? Disentangling Error Types in Entity Description Generation. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, S. 8322–8340

[Irvin u. a. 2019] IRVIN, Jeremy ; RAJPURKAR, Pranav ; KO, Michael ; YU, Yifan ; CIUREA-ILCUS, Silviana ; CHUTE, Chris ; MARKLUND, Henrik ; HAGHGOO, Behzad ; BALL, Robyn ; SHPANSKAYA, Katie u. a.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI conference on artificial intelligence* Bd. 33, 2019, S. 590–597

[Jain u. a. 2021] JAIN, Saahil ; AGRAWAL, Ashwin ; SAPORTA, Adriel ; TRUONG, Steven Q. ; DUONG, Du N. ; BUI, Tan ; CHAMBON, Pierre ; ZHANG, Yuhao ; LUNGREN, Matthew P. ; NG, Andrew Y. u. a.: Radgraph: Extracting clinical entities and relations from radiology reports. In: *arXiv preprint arXiv:2106.14463* (2021)

[Ji u. a. 2023] JI, Ziwei ; LEE, Nayeon ; FRIESKE, Rita ; YU, Tiezheng ; SU, Dan ; XU, Yan ; ISHII, Etsuko ; BANG, Ye J. ; MADOTTO, Andrea ; FUNG, Pascale: Survey of hallucination in natural language generation. In: *ACM Computing Surveys* 55 (2023), Nr. 12, S. 1–38

[Johnson u. a. 2020] JOHNSON, Alistair ; BULGARELLI, Lucas ; POLLARD, Tom ; HORNG, Steven ; CELI, Leo A. ; MARK, Roger: Mimic-iv. In: *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)* (2020)

[Johnson u. a. 2019] JOHNSON, Alistair E. ; POLLARD, Tom J. ; GREENBAUM, Nathaniel R. ; LUNGREN, Matthew P. ; DENG, Chih-ying ; PENG, Yifan ; LU, Zhiyong ; MARK, Roger G. ; BERKOWITZ, Seth J. ; HORNG, Steven: MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. In: *arXiv preprint arXiv:1901.07042* (2019)

[Kang und Hashimoto 2020] KANG, Daniel ; HASHIMOTO, Tatsunori: Improved natural language generation via loss truncation. In: *arXiv preprint arXiv:2004.14589* (2020)

[Luo u. a. 2022] LUO, Renqian ; SUN, Liai ; XIA, Yingce ; QIN, Tao ; ZHANG, Sheng ; POON, Hoifung ; LIU, Tie-Yan: BioGPT: generative pre-trained transformer for biomedical text generation and mining. In: *Briefings in Bioinformatics* 23 (2022), Nr. 6, S. bbac409

[Maynez u. a. 2020] MAYNEZ, Joshua ; NARAYAN, Shashi ; BOHNET, Bernd ; MCDONALD, Ryan: On faithfulness and factuality in abstractive summarization. In: *arXiv preprint arXiv:2005.00661* (2020)

[OpenAI 2023] OPENAI, R: Gpt-4 technical report. arxiv 2303.08774. In: *View in Article* 2 (2023), S. 3

[Petroni u. a. 2019] PETRONI, Fabio ; ROCKTÄSCHEL, Tim ; LEWIS, Patrick ; BAKHTIN, Anton ; WU, Yuxiang ; MILLER, Alexander H. ; RIEDEL, Sebastian: Language models as knowledge bases? In: *arXiv preprint arXiv:1909.01066* (2019)

[Radford u. a. 2019] RADFORD, Alec ; WU, Jeffrey ; CHILD, Rewon ; LUAN, David ; AMODEI, Dario ; SUTSKEVER, Ilya u. a.: Language models are unsupervised multitask learners. In: *OpenAI blog* 1 (2019), Nr. 8, S. 9

[Rafailov u. a. 2023] RAFAILOV, Rafael ; SHARMA, Archit ; MITCHELL, Eric ; ERMON, Stefano ; MANNING, Christopher D. ; FINN, Chelsea: Direct preference optimization: Your language model is secretly a reward model. In: *arXiv preprint arXiv:2305.18290* (2023)

[Raffel u. a. 2020] RAFFEL, Colin ; SHAZEER, Noam ; ROBERTS, Adam ; LEE, Katherine ; NARANG, Sharan ; MATENA, Michael ; ZHOU, Yanqi ; LI, Wei ; LIU, Peter J.: Exploring the limits of transfer learning with a unified text-to-text transformer. In: *The Journal of Machine Learning Research* 21 (2020), Nr. 1, S. 5485–5551

[Shi u. a. 2023] SHI, Freda ; CHEN, Xinyun ; MISRA, Kanishka ; SCALES, Nathan ; DOHAN, David ; CHI, Ed H. ; SCHÄRLI, Nathanael ; ZHOU, Denny: Large language models can be easily distracted by irrelevant context. In: *International Conference on Machine Learning* PMLR (Veranst.), 2023, S. 31210–31227

[Shi u. a. 2021] SHI, Tian ; KENESHLOO, Yaser ; RAMAKRISHNAN, Naren ; REDDY, Chandan K.: Neural abstractive text summarization with sequence-to-sequence models. In: *ACM Transactions on Data Science* 2 (2021), Nr. 1, S. 1–37

[Stiennon u. a. 2020] STIENNON, Nisan ; OUYANG, Long ; WU, Jeffrey ; ZIEGLER, Daniel ; LOWE, Ryan ; VOSS, Chelsea ; RADFORD, Alec ; AMODEI, Dario ; CHRISTIANO, Paul F.: Learning to summarize with human feedback. In: *Advances in Neural Information Processing Systems* 33 (2020), S. 3008–3021

[Sung u. a. 2021] SUNG, Mujeen ; LEE, Jinhyuk ; YI, Sean ; JEON, Minji ; KIM, Sungdong ; KANG, Jaewoo: Can language models be biomedical knowledge bases? In: *arXiv preprint arXiv:2109.07154* (2021)

[Van Veen u. a. 2023] VAN VEEN, Dave ; VAN UDEN, Cara ; BLANKEMEIER, Louis ; DELBROUCK, Jean-Benoit ; AALI, Asad ; BLUETHGEN, Christian ; PAREEK, Anuj ; POLACIN, Malgorzata ; COLLINS, William ; AHUJA, Neera u. a.: Clinical text summarization: adapting large language models can outperform human experts. In: *arXiv preprint arXiv:2309.07430* (2023)

[Vaswani u. a. 2017] VASWANI, Ashish ; SHAZEER, Noam ; PARMAR, Niki ; USZKOREIT, Jakob ; JONES, Llion ; GOMEZ, Aidan N. ; KAISER, Łukasz ; POLOSUKHIN, Illia: Attention is all you need. In: *Advances in neural information processing systems* 30 (2017)

[Wu u. a. 2023] WU, Chaoyi ; LIN, Weixiong ; ZHANG, Xiaoman ; ZHANG, Ya ; WANG, Yanfeng ; XIE, Weidi: Pmc-llama: Towards building open-source language models for medicine. In: *arXiv preprint arXiv:2305.10415* (2023)

[Yackel und Embi 2010] YACKEL, Thomas R. ; EMBI, Peter J.: Unintended errors with EHR-based result management: a case series. In: *Journal of the American Medical Informatics Association* 17 (2010), Nr. 1, S. 104–107

[Yao u. a. 2022] YAO, Zonghai ; CAO, Yi ; YANG, Zhichao ; DESHPANDE, Vijeta ; YU, Hong: Extracting biomedical factual knowledge using pretrained language model and electronic health record context. In: *AMIA Annual Symposium Proceedings* Bd. 2022 American Medical Informatics Association (Veranst.), 2022, S. 1188

[Yuan u. a. 2023] YUAN, Hongyi ; YUAN, Zheng ; TAN, Chuanqi ; WANG, Wei ; HUANG, Songfang ; HUANG, Fei: RRHF: Rank Responses to Align Language Models with Human Feedback. In: *Thirty-seventh Conference on Neural Information Processing Systems*, 2023

[Zhang u. a. 2023] ZHANG, Muru ; PRESS, Ofir ; MERRILL, William ; LIU, Alisa ; SMITH, Noah A.: How language model hallucinations can snowball. In: *arXiv preprint arXiv:2305.13534* (2023)

[Zhao u. a. 2023] ZHAO, Yao ; JOSHI, Rishabh ; LIU, Tianqi ; KHALMAN, Misha ; SALEH, Mohammad ; LIU, Peter J.: Slic-hf: Sequence likelihood calibration with human feedback. In: *arXiv preprint arXiv:2305.10425* (2023)

[Zheng u. a. 2023] ZHENG, Lianmin ; CHIANG, Wei-Lin ; SHENG, Ying ; ZHUANG, Siyuan ; WU, Zhanghao ; ZHUANG, Yonghao ; LIN, Zi ; LI, Zhuohan ; LI, Dacheng ; XING, Eric u. a.: Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: *arXiv preprint arXiv:2306.05685* (2023)

[Ziegler u. a. 2019] ZIEGLER, Daniel M. ; STIENNON, Nisan ; WU, Jeffrey ; BROWN, Tom B. ; RADFORD, Alec ; AMODEI, Dario ; CHRISTIANO, Paul ; IRVING, Geoffrey: Fine-tuning language models from human preferences. In: *arXiv preprint arXiv:1909.08593* (2019)