Direct Clinician Preference Optimization: Clinical Text Summarization via Expert Feedback-Integrated LLMs

Stanford CS224N Custom Project

Onat Dalmaz
Department of Electrical Engineering
Stanford University
onat@stanford.edu

Tim Reinhart
Department of Computer Science
Stanford University
rtim@stanford.edu

Mike Timmerman
Department of Aeronautics and Astronautics
Stanford University
mtimmerm@stanford.edu

Abstract

Healthcare services are inseparable from clinical documentation, especially text summarization, which poses a significant challenge to clinicians. This has propelled the development of automated technologies aimed at streamlining this essential yet burdensome task for clinicians. Large Language Models (LLMs) has emerged as a promising solution, yet their application in clinical settings is hampered by limitations related to proprietary models' accessibility, costs, and reliance on in-context learning. Addressing these challenges, this project introduces a novel approach that leverages an open-source, lighter-weight LLM aligned through Direct Preference Optimization (DPO) and Supervised Fine-Tuning (SFT). Our method leverages clinician feedback to learn from their preferences to tailor the model's outputs, ensuring they are aligned with the needs of clinical practice. Our experiments demonstrate that this approach improves upon the performance of the model post-SFT. Furthermore, our approach also outperforms existing open-source solutions on Open-i dataset, offering a viable alternative to proprietary models by bridging the gap between technical capabilities and clinical expectations. Our findings underscore the potential of integrating direct clinician input into LLM training processes, paving the way for more accurate, relevant, and accessible tools for clinical text summarization.

1 Introduction

Healthcare and documentation are inseparable from each other. Doctors often navigate through extensive textual information to summarize radiology reports, write progress notes, or synthesize patient histories across specializations (Golob Jr et al., 2016; Arndt et al., 2017). Reviewing and summarizing extensive textual data from electronic health records poses a significant burden to clinicians in terms of time and effort. Moreover, clinical text summarization, while critical, is an intricate and error-prone task, with inaccuracies potentially leading to serious consequences (Bowman, 2013; Gershanik et al., 2011).

The advent of Large Language Models (LLMs) represents a frontier of innovation, achieving state-of-the-art in a wide variety of natural language tasks (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020; Touvron et al., 2023). Yet, investigating their effectiveness on a diverse range of clinical summarization tasks remains understudied, partly due to lack of evaluations on relevant clinical tasks in existing benchmarks (Zheng et al., 2023; Wornow et al., 2023).

Proprietary models have recently showcased remarkable abilities in clinical text summarization tasks, igniting optimism for their application in reducing the documentation workload for clinicians (Ma et al., 2024; Van Veen et al., 2024). Despite their prowess, such models face practical limitations due to their reliance on in-context learning, (Tang et al., 2023) and the significant barriers of cost and accessibility (Burtsev et al., 2023). On the other hand, open-source models facilitate a collaborative environment for development, allowing for extensive customization and experimentation by the global NLP and healthcare community.

Despite the progress with open-source models through supervised fine-tuning (SFT) on existing datasets (Cai et al., 2023; Hu et al., 2022b, 2021), there remains a significant gap: the integration of direct clinician feedback into the model training process. This pivotal aspect of model development has been largely unexplored, highlighting an opportunity for leveraging open-source flexibility to truly align model outputs with the intricate needs of clinical practice.

In this project, we introduced Direct Preference Optimization (DPO) (Rafailov et al., 2023) into the training of a Large Language Model (LLM) tailored for radiology report summarization. We first perform SFT a sequence-to-sequence model on an open-source labeled dataset. Then, we harness expert feedback to directly optimize based on their preferences, in contrast to existing SFT methods that rely solely on fixed labeled datasets. This approach aims to bridge the gap between clinicians' expectations and model outputs and paves the way for models that are more closely aligned with the nuanced needs of clinical practice. Through experiments and analyses, we demonstrate that DPO improves upon SFT model, and our technique surpasses the performance of existing open-source solutions in radiology report summarization, providing an effective and accessible alternative to proprietary systems.

2 Related Work

2.1 Text summarization: Early Efforts and Traditional Approaches

Clinical text summarization condenses medical documents like patient records and radiology reports into concise summaries, highlighting essential diagnostic and treatment information to aid clinicians in decision-making and enhance patient care efficiency. Initial attempts at automating general text summarization in NLP relied on rule-based systems and simpler machine learning models. Among the pioneering works, LEXRANK (Erkan and Radev, 2004) stands out as a notable example of an unsupervised approach to text summarization based on graph-based centrality scoring of sentences. Although it laid foundational principles for text summarization, its application to clinical texts showed limitations in handling the domain's complexity and variability.

2.2 Advancements with Deep Learning

Earlier studies exclusively focused on sequence-to-sequence (seq2seq) methods, characterized by early use of bi-directional LSTMs, which, despite achieving some success, faced challenges in maintaining factual accuracy, with studies like (Zhang et al., 2018) noting that 30% of generated summaries contained factual errors. Innovations continued with efforts such as (MacAvaney et al., 2019), which enhanced content selection by incorporating domain-specific ontology information, and (Sotudeh Gharebagh et al., 2020), which integrated salient clinical terms using a separate encoder to refine summaries further.

2.3 Transformers and Pretraining

The rise of pre-trained language models such as BERT (Devlin et al., 2019) and GPT (Brown et al., 2020) has revolutionized medical text summarization, with approaches like TransABS (Liu and Lapata, 2019) and BertSUM (Liu, 2019) utilizing two-stage fine-tuning to achieve unparalleled results across various datasets, while CAVC method (Song et al., 2019) enhances performance through Mask Language Modeling (MLM). Simultaneously, the innovation of graph-based models and contrastive learning, as demonstrated by WGSUM and Jinpeng et al., introduces graph encoders and contrastive techniques to improve key word extraction and summary accuracy. Recently, ChestXRayBERT's domain-specific pre-training, utilizing a radiology-related corpus combined with a Transformer decoder, signified a leap in diagnostic report summarization, showcasing the effectiveness of tailoring NLP techniques to specific medical domains. Diverging from these existing open-source methodologies, our approach begins with a larger pre-trained and instruction-fine-tuned model, which we first directly adapt to the task through SFT. Followingly, we achieve further refinement of the LLM through preference learning, a novel strategy that directly integrates clinician feedback via DPO algorithm.

2.4 Very Large Proprietary Language Models

The advent of Large Language Models (LLMs) like GPT-3.5 and -4 has significantly advanced the performance automated summarization techniques. ImpressionGPT (Ma et al., 2024) attempted to adapt these very large models by means of in-context learning (ICL (Lampinen et al., 2022)). ICL embeds contextually relevant examples into the model's prompt to dynamically adapt to specific requirements, thus enhancing its understanding and summarization capabilities. Similarly, Van Veen et al. (2024) employs ICL to fine-tune its response, utilizing a select number of in-context examples. These approaches underscored the transformative potential of LLMs in clinical settings by achieving impressive results, demonstrating the value of tailored, contextual information in generating precise summaries. However, despite their advancements, these models face challenges related to their proprietary nature, substantial data requirements, and limited domain-specific generalization. Our solution seeks to address these limitations by introducing an open-source model that not only allows for the actual tuning of its weights but is also lighter-weight and specifically designed to incorporate clinician feedback through Direct Preference Optimization (DPO). Unlike the proprietary models, our approach offers a more adaptable and flexible framework that can be finely tuned to meet the intricate demands of clinical practice.

3 Approach

3.1 Large Sequence-to-Sequence Models

We focused on sequence-to-sequence (seq2seq) models due to their proven efficacy in tasks requiring nuanced language generation, such as machine translation and summarization (Raffel et al., 2019). These models, leveraging an encoder-decoder architecture, are particularly adept at mapping complex input texts to coherent, concise outputs, a capability essential for summarizing detailed radiology reports. Among the diverse collection of seq2seq models, the need for a robust and versatile architecture leads us to prioritize models pre-trained on extensive corpora, ensuring they possess a broad understanding of natural language. Consequently, we deployed the T5 (Text-to-Text Transfer Transformer) architecture (Raffel et al., 2019), specifically the FLAN-T5-XL variant, renowned for its comprehensive pretraining on the Colossal Clean Crawled Corpus (C4) (Dodge et al., 2021).

3.2 Quantized Low-Rank Adaptation Method

For efficient alignment of our pretrained LLM, we utilized the Low-Rank Adaptation (LoRA) method, which enhances fine-tuning by inserting trainable matrices (adapters) into the model architecture, thus preserving the entire sequence for downstream tasks and requiring only a minimal adjustment of total parameters (Hu et al., 2022a). Further efficiency was achieved through Quantized Low-Rank Adaptation (QLoRA), which employs 4-bit quantization to reduce memory demands significantly while maintaining model performance, allowing for the fine-tuning of larger models under hardware constraints (Dettmers et al., 2023). Throughout our subsequent alignment processes SFT and DPO, we kept these adapters trainable, without inserting new ones between the two learning paradigms.

3.3 Supervised Fine-Tuning (SFT)

As the initial step in aligning our pretrained language model, we performed Supervised Fine-Tuning (SFT). This process is pivotal for refining the model's inherent linguistic capabilities, tailoring them to produce concise and accurate clinical summaries. During SFT, we employed a standard cross-entropy loss function, which is widely utilized in sequence-to-sequence learning tasks:

$$L(\theta) = -\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} \log(p(y_{ij}|x_i;\theta)) \tag{1}$$

Here, $L(\theta)$ represents the loss for parameters θ, with N denoting the number of samples and M indicating the sequence length. Here, y_{ij} is the actual token in the generated summary, x_i is the input clinical report, and $p(y_{ij}|x_i;\theta)$ signifies the predicted probability of the token y_{ij} given the input report x_i, parameterized by θ.

3.4 Alignment via Clinician Preference Learning

To further optimize our SFTed LLM in line with clinician preferences, we leveraged Direct Preference Optimization (DPO) (Rafailov et al., 2023). This approach reformulates the traditional reinforcement learning (RL) framework, focusing on a policy objective that aligns with clinician-chosen outcomes over those less favored. Our objective is to directly maximize the likelihood of clinician-preferred summaries using the optimal policy π_θ instead of a separately learned reward model. This is encapsulated in the DPO objective function as follows:

$$\mathcal{L}_{DPO}(\pi_\theta;\pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \tag{2}$$

where σ represents the sigmoid function, β is a scaling factor that adjusts the steepness of the preference curve, (x, y_w, y_l) denotes the tuple of input text, winning summary, and losing summary sampled from the dataset $\mathcal{D}$, and π_{ref} stands for the reference policy which is calculated via a frozen reference model that has the same parameters as the initial LLM. The predicted probabilities, $\pi_\theta(y_w|x)$ and $\pi_\theta(y_l|x)$, reflect the model's preference for generating the winning summary y_w over the losing summary y_l given the input x. This loss function encourages the model to prefer clinician-selected summaries, effectively tuning the model to generate outputs that reflect clinical expertise and preferences.

4 Experiments

4.1 Data

4.1.1 SFT and evaluations

We utilized Open-i dataset's (Demner-Fushman et al., 2012) training split for SFT, dev set for hyperparameter and model tuning, and test split for evaluations. The dataset is sourced from the Indiana Network for Patient Care and consists of de-identified narrative chest x-ray reports. Originally containing 4K studies, Demner-Fushman et al. (2012) refined it to a subset of 3.4K report-summary pairs, selected for the quality of imaging views and diagnostic content.

4.1.2 DPO dataset

To capture clinician preferences for summaries within a dataset suitable for DPO, we first processed the results of an in vivo clinical reader study (Van Veen et al., 2024). To do this, we aggregated individual scores for each summary to determine its overall favorability among clinicians. Summaries were classified based on their aggregate scores, with those receiving positive scores deemed preferred or 'winning' and those with negative scores marked as less favored or 'losing'. We then structured the dataset to align each summary with its corresponding clinician preference, creating pairs of winning and losing responses for each unique input (For an example result from reader study and its corresponding structured sample, please see Tables A1 and A2, respectively).

4.2 Evaluation method

In our evaluation methodology, we employed a comprehensive set of metrics to assess the quality of the generated summaries. BLEU (Papineni et al., 2002) measures syntactic similarity through n-gram overlap, while ROUGE-1 and ROUGE-2, and ROUGE-L (Lin, 2004) evaluates the unigram, bigram, and longest common subsequence overlap, respectively. To specifically address the factual accuracy critical in the medical domain, we also incorporated F1Radgraph (Delbrouck et al., 2022) and F1chexbert (Smit et al., 2020) metrics. These two metrics, based on pretrained models, focuses on factual correctness by leveraging semantic rewards from annotated entities for accurate assessment of radiology report summaries. In our vivo experiments in 4.4.1 and 4.4.3, we reported all of the above-defined metrics, whereas in comparison 4.4.2 with other methods we only utilized ROUGE-1, ROUGE-2, and ROUGE-L as commonly done in the respective studies.

4.3 Experimental details

PyTorch, Transformers[1], TRL[2], and PEFT[3] libraries are used extensively in our implementations. Table 1 summarizes the hyperparameters used, as well as training times for SFT and DPO training procedures.

Table 1: Hyperparameters used in our experiments for QLoRA, SFT, and DPO

	Task Type	Quantization	r (rank of adaptation)	α (scale)	Dropout
QLoRA	seq2seq	4 bit	8	32	0.1

	Learning Rate	Batch Size	Beta	Loss	Number of Steps	Optimizer	Training time
SFT	1e-3	6	-	Cross-entropy	5 epochs	AdamW	7710s
DPO	1e-5	4	0.1	Log-sigmoid	500 steps	RMSProp	123s

4.4 Results

4.4.1 Benefits of Direct Preference Optimization (DPO)

To quantitatively assess the impact of DPO, we compared the performance of the models after SFT (w/o DPO) and after DPO. Results of this experiment are displayed in Figure 1, which clearly demonstrate the superior performance of our model in all metrics when augmented with DPO. Besides tailoring the summarization process more closely to clinician needs, quantitative results are also indicative of DPO's capacity to elevate the overall quality and relevance of the generated summaries. This alignment can be especially critical in clinical settings, where precision and conciseness in report summarizations can significantly impact diagnostic and treatment decisions.

4.4.2 Comparison Against State-of-the-art

In this section, we present a quantitative comparison of our models against the state-of-the-art proprietary and open-source baselines for radiology report summarization on the Open-i dataset, as illustrated in Table 2. Here we reported ROUGE-1, ROUGE-2, and ROUGE-L metrics as done in the previous studies, which are widely recognized for assessing performance in clinical text summarization studies.

Proprietary models, particularly those developed on iterations of GPT, serve as the current gold standard in the field. Particularly, the method by (Van Veen et al., 2024) based on in-context learning (ICL) via GPT-4 represents the apex of performance among proprietary solutions, which is followed by ImpressionGPT method based on a similar approach albeit leveraging GPT-3.5. Notably, our approach impressively surpasses the performance of ImpressionGPT, a proprietary model with which had been considered a benchmark. We would like to note that ImpressionGPT operates with a number of parameters that is orders of magnitude larger than those employed in our approach.

Similarly, among open-source models, our proposed methods demonstrate notable superiority. The variant incorporating the DPO framework significantly outperforms all other open-source competitors, achieving the highest scores in all metrics. This not only showcases its efficacy in summarizing radiological findings accurately but also renders it a promising and light-weight open-source alternative to proprietary models.

[1] https://huggingface.co/docs/transformers/en/index

[2] https://huggingface.co/docs/trl/en/index

[3] https://huggingface.co/docs/transformers/main/en/peft

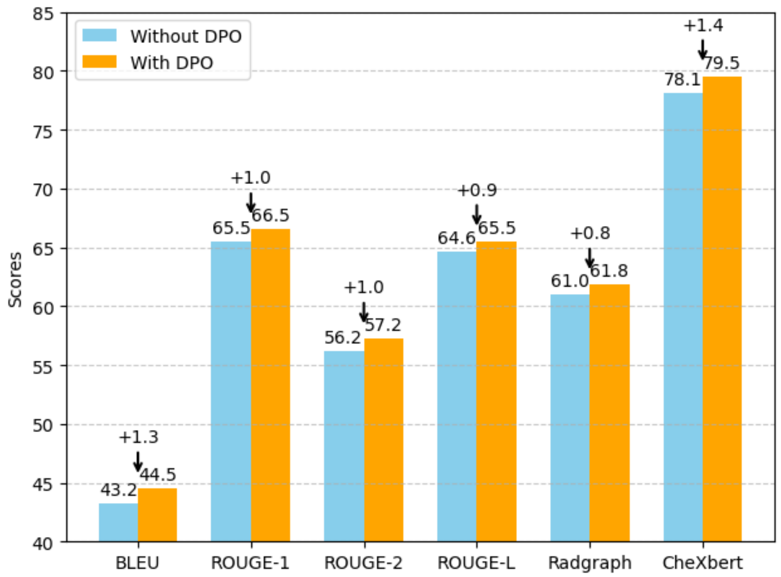

Figure 1: Improvements in model performance achieved via DPO). This bar plot compares the model's scores across various metrics, both with and without DPO, highlighting the effect of integrating clinician feedback.

Table 2: Quantitative Comparison of Baseline Models for Radiology Report Summarization on the Open-i Dataset. Proprietary models are regarded as the golden standard in the field, and the best performing open-source model is highlighted in boldface. Note that the results reported in this table are adopted from the respective papers.

Proprietary Baselines			
Model	ROUGE-1	ROUGE-2	ROUGE-L
GPT-4 ICL Van Veen et al. (2024)	-	-	68.2
GPT-3.5 ICL (ImpressionGPT) (Ma et al., 2024)	66.3	54.9	65.4
Open-Source Baselines			
Ours	**66.5**	**57.2**	**65.5**
Ours w/o DPO	65.5	56.2	64.6
Jinpeng et al. (Hu et al., 2022b)	64.97	55.59	64.45
WGSUM (LSTM) (Hu et al., 2021)	64.32	55.48	63.97
WGSum (Trans) (Hu et al., 2021)	61.63	50.98	61.73
CAVC (Song et al., 2019)	53.18	39.59	52.86
Transabs (Song et al., 2019)	59.66	49.41	59.18
ChestXRayBERT (Cai et al., 2023)	41.3	28.6	41.5
LEXRANK (Erkan and Radev, 2004)	14.6	4.4	14.1

4.4.3 Effect of DPO losses

Throughout our experiments, we have used sigmoid loss on the normalized likelihood to fit a logistic regression, based on the Bradley-Terry model (Bradley and Terry, 1952). To investigate how the type of loss function used affects learning preferences, we conducted an ablation study. To this end, we compared our vanilla log-sigmoid function with alternative approaches, such as RSO (Liu et al., 2024) based on hinge loss; IPO (Azar et al., 2023) that averages over log-likelihoods of completions through the beta parameter; KTO (Ethayarajh et al., 2024) that directly maximizes the utility of generations instead of the log-likelihood of preferences. Note that variations among the alternative loss functions—RSO, IPO, and KTO—highlight trade-offs between addressing overfitting and optimizing performance, with each exhibiting strengths in different aspects of preference learning.

Selected vanilla DPO loss function mostly outperforms alternative approaches (RSO, IPO, KTO) across metrics, suggesting its superior capability in capturing preference over other alternatives (Table 3).

Table 3: Ablation Study on Loss Functions for Preference Learning

Loss Function	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	Radgraph	CheXbert
Vanilla DPO	**44.5**	**66.5**	**57.2**	**65.5**	61.8	**79.5**
RSO(Liu et al., 2024)	44.2	66.4	57.0	**65.5**	**61.9**	79.0
IPO (Azar et al., 2023)	44.2	66.1	56.9	65.1	61.7	78.7
KTO (Ethayarajh et al., 2024)	43.6	65.5	56.6	64.6	61.3	78.5

5 Analysis

5.1 Radiology Report Summarization Outputs

To analyze and identify which model offers summaries with greater clinical relevance, we examined the outputs of DPO and SFT models in two specific example scenarios. Summaries generated by the two models for two scenarios, along with the input report and ground-truth target summaries are provided in Table 4a and 4b.

Case 1: Right Upper Lobe Pneumonia The target summary highlighted the need for a follow-up for "Right upper lobe pneumonia." The DPO model suggested opacities potentially indicative of atelectasis but misplaced them to the left side. The SFT model, on the other hand, lacked actionable insights for clinical follow-up, similarly misplacing the pathology. Despite the DPO model's mislocalization, its effort to interpret pathological findings offers slightly more clinical utility by attempting a diagnosis.

Case 2: Streaky Left Basilar Airspace Opacities In this case, the target summary considered the opacities as possibly indicative of atelectasis or infection. The DPO model specifically suggested atelectasis as a cause, closely matching one part of the target summary but missing the broader differential diagnosis. Conversely, the SFT model described the opacities as "stable," offering a less informative summary that does not guide treatment or follow-up actions.

These qualitative analyses suggests that DPO can effectively align model outputs with the needs of clinical practice, as evidenced by its tendency to generate summaries that not only describe findings but also suggest potential diagnoses and follow-up actions. This characteristic is not trivially observable in outputs from SFT model, which tend to provide less actionable information. DPO's training strategy, which hinges on clinician-selected preferences, inherently steers the model toward a decision-oriented summarization style. This focus on actionable insights could be particularly valuable in clinical settings where the quality of information synthesis directly influences medical decision-making, underscoring DPO's potential to enhance the utility of automated summarization tools in healthcare.

6 Conclusion

By aligning LLMs with fine-tuning methods like Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO), our approach not only matches with but also surpasses current open-source benchmarks. We used a sequence-to-sequence model, equipped efficient model adaptation techniques such as QLoRA. Our introduction of Direct Preference Optimization (DPO) as an integral component of the model training process has resulted in notable improvements in the generation of summaries that are both precise and clinically relevant. Empirical results highlighted our model's superior performance across key metrics. Moreover, our findings emphasize the potential of open-source models, which offer the healthcare community an opportunity for extensive customization and collaboration, thereby accelerating innovation and reducing barriers to access.

Future work in enhancing clinical text summarization models will benefit from scaling up the model architecture sizes to that of state-of-the-art, potentially increasing the model's performance. Complementing these advancements, the efficacy of DPO-tuned models should be rigorously validated

Table 4: Comparison of Radiology Report Summaries Across Different Cases

(a) Case 1: Right Upper Lobe Pneumonia

Report	The patient is rotated to left. The cardiomediastinal silhouette is normal in size. XXXX lucency along the left ventricular XXXX XXXX related to interface between the heart and aerated lung. Patchy right perihilar/upper lobe opacities, which abut the XXXX fissure on lateral projection. No pneumothorax or large pleural effusion. Exaggerated thoracic kyphosis. No definite acute bone abnormality.
Target Summary	Right upper lobe pneumonia. Consideration may be given for followup chest x-XXXX, following appropriate therapy.
DPO Model Output	1. Left basilar opacities, **possibly related to atelectasis**. 2. Patchy right perihilar/upper lobe opacities, which abut the XXXX fissure on lateral projection.
SFT Model Output	1. XXXX lucency along the left ventricular XXXX **related to interface between the heart and aerated lung.** 2. Patchy right perihilar/upper lobe opacities, which abut the XXXX fissure on lateral projection.

(b) Case 2: Streaky Left Basilar Airspace Opacities

Report	Cardiomediastinal contour and pulmonary vascularity stable and within normal limits. Lung volumes are slightly low. There are streaky left basal opacities. No pleural effusion or pneumothorax. No acute osseous findings. No free air is demonstrated.
Target Summary	Streaky left basilar airspace opacities, **which could reflect atelectasis** and/or infection.
DPO Model Output	Streaky left basilar opacities, **suggestive of atelectasis.**
SFT Model Output	1. Stable left basilar opacities.

through clinician-led reader studies, ensuring that model outputs align with the practical demands of medical practice. Furthermore, data augmentation techniques such as leveraging seed models or incorporating findings from advanced proprietary models could provide a broader and more enriched dataset for training. Such strategies will be pivotal in refining the models to better assist healthcare professionals and improve patient care.

7 Acknowledgements

- Mentor: David Lim
- External Collaborator: Akshay Chaudhari

Table 5: Team Member Individual Contributions

Team Member	Contributions
Onat Dalmaz	Training of model using DPO and creation of DPO dataset.
Tim Reinhart	Analysis, evaluation and comparison of the trained models.
Mike Timmerman	Reproducing of SFTed model and DPO loss ablation study.

References

Brian G Arndt, John W Beasley, Michael D Watkinson, Jonathan L Temte, Wen-Jan Tuan, Christine A Sinsky, and Valerie J Gilchrist. 2017. Tethered to the ehr: primary care physician workload assessment using ehr event log data and time-motion observations. *The Annals of Family Medicine*, 15(5):419–426.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences.

S Bowman. 2013. Impact of electronic health record systems on information integrity: quality and safety implications. *Perspectives in Health Information Management*, 10.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mikhail Burtsev, Martin Reeves, and Adam Job. 2023. The working limitations of large language models. *MIT Sloan Management Review*.

Xiaoyan Cai, Sen Liu, Junwei Han, Libin Yang, Zhenguo Liu, and Tianming Liu. 2023. Chestxraybert: A pretrained language model for chest radiology report summarization. *Trans. Multi.*, 25:845–855.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dina Demner-Fushman, Sameer Kiran Antani, Matthew S. Simpson, and George R. Thoma. 2012. Design and development of a multimodal biomedical information retrieval system. *J. Comput. Sci. Eng.*, 6:168–177.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

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.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization.

Esteban F Gershanik, Ronilda Lacson, and Ramin Khorasani. 2011. Critical finding capture in the impression section of radiology reports. In *AMIA Annual Symposium Proceedings*, pages 465–469. American Medical Informatics Association.

John F Golob Jr, John J Como, and Jeffrey A Claridge. 2016. The painful truth: The documentation burden of a trauma surgeon. *Journal of Trauma and Acute Care Surgery*, 80(4):742–747.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jinpeng Hu, Jianling Li, Zhihong Chen, Yaling Shen, Yan Song, Xiang Wan, and Tsung-Hui Chang. 2021. Word graph guided summarization for radiology findings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4980–4990, Online. Association for Computational Linguistics.

Jinpeng Hu, Zhuo Li, Zhihong Chen, Zhen Li, Xiang Wan, and Tsung-Hui Chang. 2022b. Graph enhanced contrastive learning for radiology findings summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J. Liu, and Jialu Liu. 2024. Statistical rejection sampling improves preference optimization.

Yang Liu. 2019. Fine-tune bert for extractive summarization.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Chong Ma, Zihao Wu, Jiaqi Wang, Shaochen Xu, Yaonai Wei, Zhengliang Liu, Fang Zeng, Xi Jiang, Lei Guo, Xiaoyan Cai, Shu Zhang, Tuo Zhang, Dajiang Zhu, Dinggang Shen, Tianming Liu, and Xiang Li. 2024. An iterative optimizing framework for radiology report summarization with chatgpt. *IEEE Transactions on Artificial Intelligence*, pages 1–12.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W. Filice. 2019. Ontology-aware clinical abstractive summarization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1013–1016, New York, NY, USA. Association for Computing Machinery.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

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

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

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.

Kaiqiang Song, Bingqing Wang, Z. Feng, Liu Ren, and Fei Liu. 2019. Controlling the amount of verbatim copying in abstractive summarization. In *AAAI Conference on Artificial Intelligence*.

Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross Filice. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1899–1905, Online. Association for Computational Linguistics.

Liyan Tang, Z. Sun, Betina R.S. Idnay, Jordan G. Nestor, Amani Soroush, Pablo Adolfo Elias, Z. P. Xu, Y. Ding, Greg Durrett, Justin F. Rousseau, Cindy Weng, and Yalei Peng. 2023. Evaluating large language models on medical evidence summarization. *NPJ Digital Medicine*, 6.

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.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Max Wornow, Yuhao Xu, Roshan Thapa, Bhavik N. Patel, Evan Steinberg, Sophia Fleming, M. A. Pfeffer, Jason Fries, and Nigam H. Shah. 2023. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6:135.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to summarize radiology findings. In *EMNLP 2018 Workshop on Health Text Mining and Information Analysis*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Appendix

A.1 DPO dataset structure

Instruction	**Output**	**Target**	**Scores from Readers**		
			q1	**q2**	**q3**
5=there is no intracranial hemorrhage...	5=1. No acute intracranial hemorrhage...	5=1. stable size of ventricles...	1	1	1
			1	0	1
			-1	0	-2
			2	1	0
			2	2	2

Table A1: A sample from the reader study and corresponding clinician scores reflecting their preference of output over target. For each unique report, scores reflect the evaluated dimensions of completeness (q1), correctness (q2), and conciseness (q3) from five different readers. The instruction column provides the input radiology report; the output is the model's generated summary; and the target is the reference summary provided by medical experts.

Report	**Winning Summary**	**Losing Summary**
there is no intracranial hemorrhage...	1. No acute intracranial hemorrhage...	1. stable size of ventricles...

Table A2: A sample from the processed dataset used for training the reward model. The ’Instruction’ column contains the original clinical text that needs summarization. The ’Winning Summary’ column represents the summary that aligns with clinician preferences, while the ’Losing Summary’ column represents the alternative summary not preferred by the clinicians, based on the aggregate scores.