shs-nlp at RadSum23: Domain-Adaptive Pre-training of Instruction-tuned LLMs for Radiology Report Impression Generation

Sanjeev Kumar Karn[*1], **Rikhiya Ghosh**[*1], **Kusuma P**[*1] **and Oladimeji Farri**[1]

[1]Digital Technology and Innovation, Siemens Healthineers
{sanjeev.kumar_karn,rikhiya.ghosh,kusuma.p,oladimeji.farri}@siemens-healthineers.com

Abstract

Instruction-tuned generative large language models (LLMs), such as ChatGPT and Bloomz, possess excellent generalization abilities. However, they face limitations in understanding radiology reports, particularly when generating the IMPRESSIONS section from the FINDINGS section. These models tend to produce either verbose or incomplete IMPRESSIONS, mainly due to insufficient exposure to medical text data during training. We present a system that leverages large-scale medical text data for domain-adaptive pre-training of instruction-tuned LLMs, enhancing their medical knowledge and performance on specific medical tasks. We demonstrate that this system performs better in a zero-shot setting compared to several *pretrain-and-finetune* adaptation methods on the IMPRESSIONS generation task. Furthermore, it ranks 1st among participating systems in Task 1B: Radiology Report Summarization at the BioNLP 2023 workshop.

1 Introduction

A radiology report is the primary method for radiologists to communicate medical image interpretations (e.g., X-rays) and conclusions to ordering physicians. These reports typically include several sections (Kahn Jr et al., 2009), with the most important ones being the FINDINGS and IMPRESSIONS sections. The FINDINGS section describes abnormalities and diagnoses, while the IMPRESSIONS section summarizes the findings and highlights major abnormalities and recommendations. A sample report with these sections can be found in Table 1.

The field of radiology has experienced rapid growth in NLP techniques due to the need for efficient and accurate analysis of radiological reports, which often consist of unstructured textual data (Ghosh et al., 2023). For instance, various efforts have been made to automatically generate IMPRESSIONS from FINDINGS, such as those by Zhang et al. (2018) using neural seq2seq, Karn et al. (2022) employing multi-agent reinforcement learning, and Delbrouck et al. (2023) utilizing Pre-trained language models (PLMs) with BERT-based systems.

FINDINGS
bifrontal hemorrhagic contusions are once again noted, stable compared to most recent prior, ... subarachnoid hemorrhage is once again noted within the left .... subdural hematoma is noted overlying the left temporal lobe and to the left .... there is no shift of normally midline structures. the ventricles appear unremarkable. a left temporal lobe hemorrhagic contusion remains stable in size .... the visualized paranasal sinuses are clear. there is no evidence of acute fracture.
IMPRESSIONS
1. bifrontal hemorrhagic contusion appears stable compared to most recent prior with slightly increased vasogenic ...
2. subdural hematoma is noted layering over the left temporal lobe and within the left falx.
3. subarachnoid hemorrhage is noted within the left frontal region.
4. no shift of normally midline structures.

Table 1: FINDINGS (top) and IMPRESSIONS (bottom) sections of a radiologist's report from MIMIC-III.

PLMs are trained on vast, diverse corpora, allowing them to capture various linguistic patterns (Gururangan et al., 2020). Despite their strengths, they have limitations and biases affecting out-of-domain tasks. Domain-specific PLMs, like BioBERT (Lee et al., 2020) and RadBERT (Yan et al., 2022), have been developed for biomedical and clinical NLP. However, these models face constraints, such as RadBERT's training on small datasets with limited anatomical specialties.

The *pretrain-and-finetune* paradigm for PLMs has emerged as the prevailing approach to address downstream tasks with substantial training data scarcity (Karn et al., 2021). Recent studies (e.g., Gururangan et al. (2020)) suggest that additional pretraining on in-domain text, known as specialist pretraining, is more effective in enhancing downstream performance. Currently, another approach

[*] Equal Contribution.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 550–556

has surfaced, where instead of finetuning PLMs to perform downstream tasks, the objectives of downstream tasks are reformulated using textual prompts similar to the original pre-training objectives (Liu et al., 2023). This *pretrain-and-prompt-tune* paradigm is commonly referred to as prompt-tuning.

Multitask prompted finetuning (also known as instruction tuning) is a type of large-scale *pretrain-and-prompt-tune* paradigm in which large PLMs (also referred to as LLMs) are finetuned using datasets representing various NLP tasks, defined by instructions as natural language prompts (Scao et al., 2022). Utilizing this approach, Scao et al. (2022) endowed their LLM, Bloom, with the ability to perform multilingual zero-shot instruction-based tasks and labeled it Bloomz.

We propose an extension to the domain adaptation paradigm beyond the typical method of *pretrain-and-finetune* for domain-specific tasks. We posit that further pretraining of instruction-tuned LLMs, already subjected to the *pretrain-and-prompt-tune* process, with in-domain text will improve and simplify adaptation. We refer to our approach as *general-pretrain-prompt-tune-and-special-pretrain*. In this approach, the model is trained using the same initial LM objective in each of the three training stages (i.e., general pretraining, prompt-tuning, and domain-specialized pretraining), which is a significant advantage.

We continued pretraining the instruction-tuned Bloomz on MIMIC-IV to form RadBloomz and evaluated this adaptation paradigm on the radiology report summarization task. The proposed system, in a zero-shot setting, demonstrates better performance than *pretrain-and-finetune* methods and ranks 1st among participating systems in Task 1B: Radiology Report Summarization at the BioNLP 2023 workshop.

Overall, our contributions are as follows:

- We expand the domain adaptation paradigm by introducing *general-pretrain-prompt-tune-and-special-pretrain*, which further pretrains instruction-tuned LLMs, such as Bloomz, on domain-specific text.
- We demonstrate that the new adaptation paradigm for an instruction-tuned LLM in radiology yields better performance in a zero-shot setting compared to the *pretrain-and-finetune* methods.

Dataset	FINDINGS	IMPRESSIONS
MIMIC-IV	113.46(139.06)	33.04 (36.52)
MIMIC-III	118.19(59.7)	49.48 (35.12)
MIMIC-CXR	54.52 (24.67)	16.37 (15.79)

Table 2: Number of words per report with standard deviation in parentheses for various dataset.

2 Datasets

2.1 Pretraining Datasets for Radiology Domain Adaptation.

We performed domain-adaptive pretraining using the recently published MIMIC-IV radiology reports dataset (Johnson et al.; PhysioBank, 2000). This dataset contains over 2.3 million radiology reports from 237k patients and amounts to approximately 616 million tokens using the Bloomz tokenizer (Muennighoff et al., 2022). After preprocessing, we utilized only 1.4 million reports with 190 million tokens. Further details on the dataset statistics can be found in Table 2.

2.2 Finetuning Datasets for Impression Generation

We utilized the datasets shared for Task 1B: Radiology Report Summarization at the BioNLP 2023 workshop for our fine-tuning task. This task comprises three datasets: MIMIC-III (Johnson et al., 2016), MIMIC-CXR (Johnson et al., 2019), and CHEXPERT (Irvin et al., 2019), pre-split into FINDINGS and IMPRESSIONS sections. For MIMIC-III, there are 59,320 reports in the training dataset, 7,413 in validation, 6,526 in the test set, and 6,531 in the hidden test set. Most reports (91.4%) pertain to CT imaging, with the most represented anatomy being the head (52.8%). Although the task related to MIMIC-CXR/CHEXPERT datasets is multimodal, we only used radiology reports for fine-tuning and inference. The MIMIC-CXR training dataset has 125,417 reports for training, 991 for validation, and 1,624 for testing. The hidden test dataset, a CHEXPERT dataset, contains 1,000 reports for evaluation.

3 Methods

Our methods comprise preprocessing, domain-adaptive pretraining, fine-tuning, and inference.

3.1 Preprocessing

In the preprocessing step, Regex-based cleaning and normalization were used to remove irrelevant

characters and texts from the report. Special tokens for de-identified text were incorporated, and distinct sections, such as FINDINGS and IMPRESSIONS, were identified. Reports with containing both sections and fewer than 512 tokens were selected.

3.2 Domain adaptive pretraining (DAPT)

We chose GPT-powered Bloom (Scao et al., 2022) as our study's base LLM. Bloom, a multilingual LLM adapted from Megatron-LM GPT2 (Shoeybi et al., 2019), has multiple versions based on parameters. The largest has 176 billion parameters, 70 layers, 112 attention heads, and 14,336-dimensional hidden layers. We use Bloomz (Muennighoff et al., 2022), a massive multitask instruction-tuned version of Bloom, specifically its Bloomz-7b1 variant with 7 billion parameters, 30 layers, and 4,096-dimensional hidden layers for domain adaptation.

Following our proposed *pretrain-fine-tune-and-pretrain* paradigm, we continuously pretrain Bloomz-7b1 using cross-entropy loss on auto-regressively generated tokens from the FINDINGS and IMPRESSIONS sections of MIMIC-IV reports.

3.3 Finetuning

In this study, the domain-specific task for finetuning an LLM is Radiology Report Summarization. Using standard prompt-based fine-tuning, we employ FINDINGS and TL;DR as prompt, and fine-tune Bloomz-7b1 by comparing auto-regressively generated summary tokens to ground-truth IMPRESSIONS using cross-entropy loss. This method ensures fine-tuning consistency with base Bloom and intermediate Bloomz's pretraining and instruction-tuning objectives. To prevent catastrophic forgetting, Bloomz-7b1's trainable parameters are minimized by only allowing the last layer to be modified.

3.4 Inference

The inference pipeline utilizes the trained model to generate IMPRESSIONS based on the given FINDINGS. Evaluation metrics for the generated results include Rouge scores (Lin, 2004), F1RadGraph (Delbrouck et al., 2022a), Bertscore (Zhang et al., 2019), and F1CheXbert (Xie et al., 2023) for the MIMIC-CXR and CHEXPERT datasets.

4 Experiments

We propose two experimental runs for the summarization task.

1. **Radiology Domain Adaptive Pretraining (RadBloomz) with MIMIC-IV and zero-shot inference.** The Bloomz-7b1 model is fine-tuned with a causal language objective on MIMIC-IV radiology reports, creating RadBloomz. With a sequence length of 512, training batch size of 64, validation batch size of 32, learning rate of 3e-5, and AdamW optimizer (Loshchilov and Hutter, 2017), the best zero-shot inference is achieved at 24k steps.

2. **RadBloomz finetuned with MIMIC-III.** Following the *pretrain-and-finetune* paradigm, RadBloomz is further fine-tuned with the MIMIC-III dataset for radiology report summarization. Using the same hyperparameters and training configuration, the best results are achieved at 2697 steps.

All experiments were conducted on the same infrastructure, utilizing eight Tesla A100 SXM4 GPUs (80GB memory each) and Deepspeed zero-3 configuration (Rasley et al., 2020) with BF16 enabled. A sampling-based technique was used to generate summaries from the model output distribution, with a maximum of 128 tokens, top_k set to 50, and top_k at 0.7.

5 Results and Discussion

We evaluated RadBloomz against other systems using ROUGE for n-gram overlap and F1RadGraph for fact overlap.

Table 3 showcases RadBloomz's (shs-nlp team) performance on MIMIC-CXR and MIMIC-III hidden test datasets. The MIMIC-III hidden dataset contains only reports, while MIMIC-CXR includes reports and images. Since our system is text-based, MIMIC-III is a more suitable evaluation. RadBloomz ranks first in MIMIC-III and fourth in MIMIC-CXR among all submitted systems, demonstrating the effectiveness of our domain adaptation technique in a multi-modal task.

In Table 4, we compare the performance of the standard *pretrain-and-finetune* approach with our proposed *pretrain-prompt-tune-pretrain-and-zero-shot* paradigm on the radiology report summariza-

Team	hidden testset	BLEU4	ROUGE-L	BertScore	F1-cheXbert	F1-RadGraph
shs-nlp	MIMIC-III	**18.36**	**35.32**	**57.26**	N/A	**36.94**
utsa-nlp		16.05	34.41	57.08	N/A	36.31
aimi		16.61	33.43	55.54	N/A	35.12
sinai		17.38	32.32	55.04	N/A	33.96
knowlab		13.23	32.02	55.64	N/A	33.39
dmis-msra	MIMIC-CXR	**18.62**	**34.57**	**55.90**	**72.36**	**43.20**
utsa-nlp		16.33	34.97	55.54	69.41	42.86
knowlab		14.41	33.63	54.72	67.20	39.98
shs-nlp		14.59	32.43	53.99	68.99	38.40
aimi		5.15	31.84	47.83	64.18	32.05

Table 3: The table presents the performance of the top-5 submitted systems on the two categories of hidden test data for Shared Task 1B at BioNLP 2023. Our RadBloomz system is represented by shs-nlp. The hidden test set MIMIC-III contains only reports, while MIMIC-CXR includes both reports and images. As our system is text-based, MIMIC-III serves as a more appropriate evaluation, and it ranks 1st among the participating systems.

Models	open test-set	BLEU4	ROUGE-L	BertScore	F1-cheXbert	F1-RadGraph
RBz-0shot	**MIMIC-III**	**17.33**	33.93	55.49	N/A	**34.93**
RBz-ft		16.49	**35.25**	**57.29**	N/A	31.12
RBz-0shot	**MIMIC-CXR**	**25.32**	**47.48**	**63.61**	**74.34**	**49.00**
RBz-ft		16.16	26.16	52.22	53.1	31.07

Table 4: Results for various domain adaptation methods on different test splits of Shared Task 1B are shown. The experimental setup remains consistent for all methods, meaning that the same train/validation/test split of the medical reports was used. RBz-0shot represents RadBloomz-zero shot, while RBz-ft stands for RadBloomz-finetuned.

tion test data for Task 1B challenge.[1] We observe that while finetuning with MIMIC-III improves Rouge-L and Bertscore metrics for the MIMIC-III test dataset, Bertscore, F1-RadGraph, and F1-cheXbert scores are lower for the finetuned model. This indicates that the domain adaptation paradigm is sufficient for achieving higher performance without requiring task-specific finetuning.

5.1 Error Analysis

A thorough error analysis on the open test datasets reveals that many generated IMPRESSIONS receive low scores for both Rouge and F1-RadGraph when the ground-truth radiology report IMPRESSIONS does not mention any abnormalities. For instance, the generated impression "normal MRI of the cervical spine" and the ground truth impression "negative study" are semantically similar. However, these n-gram overlap-based scores fail to recognize their semantic relatedness.

Similarly, we observed that similar FINDINGS sometimes generate different impressions. For example, impressions can be as detailed as: "near complete opacification of the ethmoid air cells and sphenoid sinuses, moderate air-fluid level with mucosal thickening of the right maxillary sinus, and moderate mucosal thickening of the left maxillary sinus." Meanwhile, similar findings in another report might be summarized as "pansinusitis, as described above."

The detailed error analysis has also uncovered several intriguing types of hallucinations by RadBloomz on theMIMIC-III test dataset. More details can be found in the Appendix, Section A.1.

6 Conclusion

In this work, we introduce a new domain adaptation paradigm of *general-pretrain-prompt-tune-and-special-pretrain*, where we further pretrain an instruction-tuned LLM (Bloomz) on radiology domain text. We use radiology report summarization as the domain-specific task and demonstrate that the new paradigm-based LLM model outperforms the standard "pretrain-and-finetune" method, even in a zero-shot setting. The system ranks 1st among participating systems in the hidden-test category for Task 1B: Radiology Report Summarization at the BioNLP 2023 workshop.

Limitations

There are a few limitations pertaining to the training data we used, some of which are listed below.

1. Our domain adaptation of LLMs was performed on English reports only; therefore, it may not work out of the box in a multilingual setting.

[1] Ground-truths are available only for the open test data for any additional evaluation.

2. The paper utilizes the MIMIC-IV dataset for DAPT training, which might include overlapping data from MIMIC-III and MIMIC-CXR. Consequently, there is a potential risk of information leak in this method.

3. There is a data imbalance concerning imaging modalities and anatomies covered by our training data. For example, regions such as extremities, neck, spine, and shoulder are underrepresented in the dataset, and report summarization related to those regions needs thorough evaluation.

4. A study is needed to examine the diversity of patients represented in the data and how it impacts the model's performance for underrepresented communities.

5. Different radiologists and radiology departments have distinct preferences and styles for writing reports. Moreover, clinical referrals occasionally dictate the extent to which certain details are documented in the report. No study has been conducted on the consistency, uncertainty, or information richness of the report.

Aside from the training data, the model's space and time throughputs may render them unsuitable for on-premise and/or at-the-edge applications. This aspect presents an opportunity for further research on how to best quantize and deploy RadBloomz (and similar LLMs) within the clinical workflow to enhance efficiency for radiologists. Additionally, the paper utilizes the MIMIC-IV dataset for DAPT training, which could contain overlapping data from MIMIC-III and MIMIC-CXR. Consequently, there is a potential risk of information leak in this method.

Ethics Statement

The research performed in this paper adheres to the Association for Computing Machinery (ACM) Code of Ethical and Professional Conduct [2] adopted by the Association for Computational Linguistics (ACL). To prevent any harm caused due to errors in our model-generated outputs, our models are meant to be deployed in a human-in-the-loop setting where the key information extracted by our models are reviewed by radiologists and physicians.

[2]https://www.acm.org/code-of-ethics

Disclaimer

The concepts and information presented in this paper/presentation are based on research results that are not commercially available. Future commercial availability cannot be guaranteed.

Acknowledgements

We sincerely thank our DTI team for their support, the Siemens Healthineers supercomputing team for the training infrastructure, and anonymous reviewers for valuable feedback. We also acknowledge the shared task organizers and ViLMedic (Delbrouck et al., 2022b) for their vital contributions and efforts in ensuring a smooth, productive experience for all shared task participants.

References

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis P Langlotz. 2022a. Improving the factual correctness of radiology report generation with semantic rewards. *arXiv preprint arXiv:2210.12186*.

Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 2022b. 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*, pages 23–34, Dublin, Ireland. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Maya Varma, Pierre Chambon, and Curtis Langlotz. 2023. Overview of the radsum23 shared task on multi-modal and multi-anatomical radiology report summarization. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Rikhiya Ghosh, Sanjeev Kumar Karn, Manuela Daniela Danu, Larisa Micu, Ramya Vunikili, and Oladimeji Farri. 2023. Radling: Towards efficient radiology report understanding.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv-note: Deidentified free-text clinical notes.

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

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Charles E Kahn Jr, Curtis P Langlotz, Elizabeth S Burnside, John A Carrino, David S Channin, David M Hovsepian, and Daniel L Rubin. 2009. Toward best practices in radiology reporting. *Radiology*, 252(3):852–856.

Sanjeev Kumar Karn, Francine Chen, Yan-Ying Chen, Ulli Waltinger, and Hinrich Schütze. 2021. Few-shot learning of an interleaved text summarization model by pretraining with synthetic data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 245–254, Kyiv, Ukraine. Association for Computational Linguistics.

Sanjeev Kumar Karn, Ning Liu, Hinrich Schuetze, and Oladimeji Farri. 2022. Differentiable multi-agent actor-critic for multi-step radiology report summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1542–1553, Dublin, Ireland. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning.

PhysioToolkit PhysioBank. 2000. Physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Qianqian Xie, Jinpeng Hu, Jiayu Zhou, Yifan Peng, and Fei Wang. 2023. Factreranker: Fact-guided reranker for faithful radiology report summarization. *arXiv preprint arXiv:2303.08335*.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. 2022. Radbert: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to summarize radiology findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213, Brussels, Belgium. Association for Computational Linguistics.

A Appendix

A.1 Error Analysis

A detailed error analysis on the MIMIC-III test dataset has uncovered several intriguing types of hallucinations generated by the RadBloomz model. These hallucinations can be categorized as follows:

A.1.1 Numerical Hallucination

The RadBloomz model has been observed to hallucinate, particularly when numbers are involved. It struggles with comparing numbers or associating mentioned numbers with their correct concepts. One example can be seen in Table 5, where the

model has difficulty understanding the largest size of the lymph node conglomerate mentioned in the text, even though it is explicitly stated which one is the largest conglomerate.

FINDINGS
. . . there is massive lymphadenopathy in the mesentery and retroperitoneum with significant interval worsening compared to prior scan. there are conglomerates of lymph nodes, the largest at the paraaortic region measures 6.7 x 4.4 cm, (2:73). there is a large conglomerate to the right common iliac artery, measuring 4.7 x 4.6 cm, (2:91). . . .
Generated IMPRESSIONS
interval worsening of the massive lymphadenopathy in the mesentery and retroperitoneum, with conglomerates of lymph nodes, largest at the right common iliac artery, measuring 4.7 x 4.6 cm.

Table 5: Numerical hallucination example: FINDINGS (top) and IMPRESSIONS (bottom) sections of a radiologist's report from the MIMIC-III test dataset. The sizes of lymph node conglomerates are mentioned for both the paraaortic region and the right common iliac artery in the FINDINGS. However, the model incorrectly assumes that the right common iliac artery is larger. This incorrect arithmetic determination of sizes leads to hallucination.

A.1.2 Missing Context and Facts

There are FINDINGS in which the RadBloomz model has overlooked facts or failed to understand the context of a sentence well. One such example is illustrated in Table 6.

FINDINGS
. . . there is massive inflammatory stranding surrounding the pancreas. no definite fluid collections are seen. . . . there is a moderate amount of free fluid seen in the abdomen in the left paracolic gutter. this demonstrates attenuation values of simple fluid. there is also some fluid around the spleen. . . .
Generated IMPRESSIONS
. . . 2. extensive inflammatory stranding surrounding the pancreas. 3. no definite fluid collections are seen. . . .

Table 6: Missing Context and Facts: The FINDINGS (top) and IMPRESSIONS (bottom) sections of a radiologist's report from the MIMIC-III test dataset. The absence of fluid collection is observed only around the pancreas, but the model generates a summary that implies there is no fluid collection in any other imaged anatomical part. However, the FINDINGS mention fluid collections in the abdomen and around the spleen, which are overlooked.

A.1.3 Hallucination of New Facts

New concepts, unrelated to what is mentioned in the findings, are a part of the summary. These occurrences are rare. We found only 6 cases out of the 6.5k where this occurred. One such example is shown in Table 7.

FINDINGS
moderate right and small left pleural effusions have increased from prior examination. there is associated compressive atelectasis. the heart is normal in size, without pericardial effusion. note is made of a small sliding hiatal hernia.
Generated IMPRESSIONS
. . . no evidence of recurrent or metastatic disease in the abdomen or pelvis. . .

Table 7: Incorrect Concept Hallucination: The FINDINGS (top) and IMPRESSIONS (bottom) sections of a radiologist's report from the MIMIC-III test dataset. The FINDINGS section does not mention the abdomen or pelvis, but the generated summary includes observations in these two anatomical regions.