NTCIR-18: Proceedings of the 18th NTCIR Conference on Evaluation ofInformation Access Technologies, June 10-13, 2025, Tokyo, Japan
DOI: https://doi.org/10.20736/0002002086

RAD-PHI3 at the NTCIR-18 HIDDEN-RAD: Hidden Causality Inclusion in Radiology Reports with Multimodal Small Language Models

Mercy Ranjit
Microsoft Research India
India
meranjit@microsoft.com

Rahul Kumar
Microsoft Research India
India
b-rahulkumar@microsoft.com

Shaury Srivastav
Microsoft Research India
India
t-ssrivastav@microsoft.com

Anirban Porya
Microsoft Research India
India
t-anporya@microsoft.com

Tanuja Ganu
Microsoft Research India
India
taganu@microsoft.com

Abstract

This paper presents the participation of the Microsoft Research RADPHI3 team in the Hidden-RAD Challenge: Hidden Causality Inclusion in Radiology Reports [7]. The task aims to recover hidden causality from radiology reports, optionally accompanied by their corresponding frontal chest X-rays (CXRs).

We fine-tune small language models, specifically Rad-Phi-3.5-Vision-CXR, to recover causality analysis in both language-only and multi-modal settings, given radiology reports and radiology images as inputs. We also include baselines of various models in the general domain, including models specifically tuned for reasoning tasks such as GPT-4o [8], LLaMA 3.3[1], Phi-4 [2], DeepSeek [4], OpenAI o1 [9], OpenAI o1-mini [9], and OpenAI o3-mini[2]. Through these experiments, we evaluated the effectiveness of general-domain, reasoning-specialized, and fine-tuned domain-specific small language models in generating causal explanations given radiology reports and images optionally as inputs.

Keywords

Causal Reasoning, Radiology Reports, Small-Language Models

Team Name

RADPHI3

Subtasks

Task 1 [7]

1 Introduction

Traditional radiology reports focus primarily on documenting the symptoms, imaging findings, and final diagnosis of the patient. However, they often lack a comprehensive explanation of the causal relationships and reasoning that lead to a specific diagnosis. This omission presents a critical gap in medical reporting, as understanding the underlying decision-making process is essential for both clinical accuracy and educational purposes. Incorporating causal reasoning into radiology reports can enhance the interpretability and reliability of diagnostic assessments, ultimately improving patient care.

The Hidden-Rad Challenge [7] seeks to address this limitation by developing methodologies that explain why a diagnosis is made during the image interpretation process. By introducing structured approaches to articulate the reasoning behind diagnoses, this challenge serves as a foundational step toward generating more accurate and meaningful medical reports. The primary goal of the challenge is to evaluate how effectively a diagnosis or the diagnostic process can be integrated with a given radiology image or medical report, particularly within the context of the MIMIC database [6]. By fostering advancements in causal understanding, this initiative aims to enhance the transparency and precision of radiological assessments, contributing to the broader field of medical AI and automated reporting.

We contribute to the Task 1 [7] of the Hidden-Rad Challenge which aim to address the lack of explicit causal reasoning within traditional reports by generating a causality exploration section for radiology diagnosis. We fine-tune small language models that can map radiology reports, optionally accompanied by frontal chest X-ray images to a causal explanation summary that reflect expert diagnostic reasoning.

Key Contributions towards the task of hidden causality inclusion in radiology reports:-

(1) Baseline measurements of various general-domain large language models, including models specialized for reasoning.
(2) Evaluation under multi-modal settings by providing frontal chest X-rays along with corresponding radiology reports for generating causal exploration summaries.
(3) Demonstration that the performance of specialized small multi-modal radiology models is on par with large general-domain models.

2 Method

2.1 Dataset

The dataset for this task is sourced from the MIMIC-CXR [6] database, which includes radiology reports and their corresponding DICOM images. The dataset contains paired examples of radiology reports and their corresponding images with expert-verified causal

[1]https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md
[2]https://openai.com/index/o3-mini-system-card/

Table 1: Experimental Settings.

Base Model	Setting	Training Data
GPT 4o	Text	NA
OpenAI o1-mini	Text	NA
OpenAI o3-mini	Text	NA
LLaMA-3.3 (70B)	Text	NA
Phi-4 (14B)	Text	NA
Deepseek (70B)	Text	NA
Rad-Phi-3.5-Vision-CXR (4.2B)	Text	Released data + augmented data
GPT 4o	Image + Text	NA
OpenAI o1	Image + Text	NA
Rad-Phi-3.5-Vision-CXR (4.2B)	Image + Text	Released data + augmented data

Table 2: Hidden-RAD Dataset.

Dataset Name	Train	Test
Hidden-RAD	1165	305

explanations, enabling models to learn how radiologists derive causal relationships from imaging findings. The dataset statistics are available in Table 2.

To enhance the size and diversity of our dataset, we employed few-shot prompting with the LLaMA 3.3 language model to augment the provided dataset with 10,000 additional samples. The prompt is available in Table 7.

2.2 Experimental Design

The experiments included hidden causality restoration in both text-only (radiology reports) and multi-modal (radiology report+images) settings using both general domain models including models specialized for reasoning and radiology specialized multi-modal small language models.

We experimented on several general-domain models, including GPT-4o, LLaMA 3.3, Phi-4, DeepSeek (70B), OpenAI o1-mini, and OpenAI o3-mini in text-only setting and GPT-4o and OpenAI o1 models in multi-modal setting incorporating frontal chest X-ray images alongside radiology reports.

We fine-tuned Rad-Phi-3.5-Vision-CXR, a multi-modal version of a small language model optimized for radiology tasks [10]. The fine-tuning process utilized a combination of released data and augmented data to enhance its domain-specific reasoning. The model was fine-tuned under two settings: one to generate causal exploration summaries from radiology reports, and another to generate them from both radiology reports and images.

The experimental settings are available in Table 1.

2.3 Model Architecture

Rad-Phi-3.5-Vision-CXR is based on Phi-3.5-vision-instruct [1] architecture. It consists of a CXR-specific image encoder, an adapter layer and a small-language model. This model was instruction-tuned on various radiology-specific datasets [10]. The image encoder was trained using a contrastive learning setup leveraging sigmoid loss for language-image pre-training [11]. The sigmoid loss function helps to reduce higher penalization towards multiple positive image text-pairs in a batch as it handles samples independently of other samples in a batch. The input image resolution used is 336×336.

2.4 Training Details

We fine-tuned RAD-Phi-3.5-Vision-CXR model with the image encoder frozen using low rank adaptation method [5] in both image + text and text-only settings. In both configurations, we employed a batch size of 32 and trained the model for 3 epochs with a learning rate of 5e-5. For both the experiments, we utilized the AdamW optimizer with a cosine learning rate scheduler and a linear warm-up strategy.

2.5 Evaluation Metrics

A LLM based evaluation scheme was used to assess the accuracy and consistency of diagnostic reports based on:

- Contextual Similarity: Evaluates how semantically aligned the report is with the diagnostic and observational content of the Gold Standard.
- Consistency with Diagnostic Basis: Evaluates how accurately the report reflects the diagnostic basis outlined in the Gold Standard.
- Consistency with MIMIC Items: Evaluates how accurately the report reflects observations stated in the Gold Standard.
- Evaluation of Hidden Causality Restoration: Evaluates how clearly the report restores and reflects causality stated in the Gold Standard.

A detailed definition is available in the repository[3].

For lexical performance, BERTScore [12], Cosine Similarity, and BioSentVec [3] were used. Additionally, a qualitative assessment by radiologists was performed for further validation. The weightage assigned to these metrics is provided in Table 6.

3 Results

In the text-only setting involving radiology reports, the radiology-specific small-language model Rad-Phi-3.5-Vision-CXR achieved the best performance. Among general-domain models, GPT-4o was

[3] https://github.com/hidden-rad/Evaluation-Scheme-Experiment-

Table 3: Hidden-RAD Test Set Results - Report-only Setting.

Model Name	BERTScore	BioSentVec	COS Similarity	GPT Score White	GPT Score Black	Weighted Score
Deepseek (70B)	0.242	0.724	0.411	0.641	0.66	0.503
LLaMA-3.3	**0.27**	0.74	0.49	0.65	0.663	0.514
OpenAI o3-mini	0.238	0.754	0.472	**0.681**	0.703	0.532
OpenAI o1-mini	0.254	0.758	0.5	0.677	0.701	0.534
Phi-4	0.233	0.761	0.476	0.676	**0.71**	0.534
GPT 4o	0.238	0.765	0.495	**0.681**	0.702	0.535
Rad-Phi-3.5-Vision-CXR	0.256	**0.766**	**0.541**	0.68	0.7	**0.538**

Note: Here the weighted score does not include the qualitative evaluation score.

Table 4: Hidden-RAD Test Set Results - Report + Image Setting.

Model Name	BERTScore	BioSentVec	COS Similarity	GPT Score White	GPT Score Black	Weighted Score
OpenAI o1	0.229	0.748	0.459	0.677	0.698	0.528
GPT 4o	0.236	**0.77**	0.522	**0.691**	**0.713**	**0.543**
Rad-Phi-3.5-Vision-CXR	**0.259**	0.767	**0.538**	0.683	0.696	0.538

Note: Here the weighted score does not include the qualitative evaluation score.

Table 5: Hidden-RAD Leaderboard Rank

Model Name	Setting	BERT	BioSentVec	COS Sim	GPT White	GPT Black	Qual. Score	Weighted Score
OpenAI o1-mini	Text	0.254	0.758	0.5	0.677	0.701	–	0.671
GPT 4o	Text	0.238	0.765	0.495	0.681	0.702	–	0.672
Rad-Phi-3.5-Vision-CXR	Image + Text	**0.259**	0.767	0.538	0.683	0.696	0.682	0.675
Rad-Phi-3.5-Vision-CXR	Text	0.256	0.766	**0.541**	0.68	0.7	0.690	0.676
GPT 4o	Image + Text	0.236	**0.77**	0.522	**0.691**	**0.713**	**0.713**	**0.682**

Note: Here the weighted score includes the qualitative evaluation score.

the top performer, followed by reasoning-focused models such as OpenAI o1-mini, OpenAI o3-mini, and Phi-4, which exhibited similar performance levels. Deepseek (70B) had the lowest performance in this category.

In the multi-modal setting, involving radiology reports and frontal chest X-ray images, GPT-4o had the best performance, closely followed by Rad-Phi-3.5-Vision-CXR. Notably, GPT-4o outperformed the reasoning fine-tuned OpenAI o1 model in the multi-modal setting and demonstrated improved performance compared to the text-only setting.

In text-only settings, smaller language models, such as Rad-Phi-3.5-Vision-CXR, OpenAI o1-mini, OpenAI o3-mini, and Phi-4, either outperformed or were on par with larger models like GPT-4o, Deepseek (70B), and LLaMA 3.3. Similarly, in the multi-modal setting, Rad-Phi-3.5-Vision-CXR was very close in performance to the larger GPT-4o model.

The metrics are available in Table 3 and Table 4. The weighted score in these are computed according to Table 6 without the qualitative evaluation score.

The final leaderboard rankings of our best models for Task 1 of the Hidden-RAD Challenge, including the qualitative evaluation, are presented in Table 5. Our top-performing model is Rad-Phi-3.5-Vision-CXR, fine-tuned in the text-only settings.

4 Conclusion

In this study, we conducted a comprehensive evaluation of various general-domain models, reasoning models, and models fine-tuned on radiology-specific datasets for Task 1 of the Hidden-RAD Challenge, which is to recover hidden causality in radiology reports. Additionally, we introduced a fine-tuned version of a small multi-modal model Rad-Phi-3.5-Vision-CXR, capable of generating causality exploration from radiology reports, with the option to incorporate frontal chest X-ray images. Our results demonstrate that smaller models achieve performance on par with, in some cases exceed, that of larger language models in both general domain and domain-adapted settings and also in text-only and multi-modal settings. Future work can focus on the use of these recovered causal explorations in automated radiology reporting workflows and large scale experiments of this pilot challenge.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219 [cs.CL] https://arxiv.org/abs/2404.14219

[2] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 Technical Report. arXiv:2412.08905 [cs.CL] https://arxiv.org/abs/2412.08905

[3] Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. BioSentVec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 1–5.

[4] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] https://arxiv.org/abs/2501.12948

[5] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL] https://arxiv.org/abs/2106.09685

[6] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR Database (version 2.0.0). PhysioNet. https://doi.org/10.13026/C2JT1Q.

[7] Key-Sun Choi and Yousang Cho and Hidden-Rad Organizing Committee. 2025. Overview of the NTCIR-18 HIDDEN-RAD Task: Hidden Causality Inclusion in Radiology Report Generation. In *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies*.

[8] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini

Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. GPT-4o System Card. arXiv:2410.21276 [cs.CL] https://arxiv.org/abs/2410.21276

[9] OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. 2024. OpenAI o1 System Card. arXiv:2412.16720 [cs.AI] https://arxiv.org/abs/2412.16720

[10] Mercy Ranjit, Shaury Srivastav, and Tanuja Ganu. 2024. RadPhi-3: Small Language Models for Radiology. arXiv:2411.13604 [cs.CV] https://arxiv.org/abs/2411.13604

[11] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11975–11986.

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

A APPENDIX

Table 6: Weightage of Metrics

Metric	**Weightage (%)**
BERTScore	5
COS Similarity	5
BioSentVec	20
GPT Base Score (Black)	25
GPT Base Score (White)	25
Qualitative Score	20

Table 7: Prompt for dataset generation (LLaMA 3.3).

You are a radiology expert. The user will send a radiology report and your task is to generate causal explanation for the given radiology report.
a. You will first clean the given report as a preprocessing step:

- Eliminate irrelevant comparisons (e.g., *"Chest x-ray from ___"*).
- Remove any mention of date and time (e.g., *"The chest PA and lateral views submitted at 14:08"*).
- Retain essential anatomical observations (e.g., *"pleural effusion, patchy opacities, diaphragm elevation"*).
- Include important imaging abnormalities (e.g., *"right costophrenic angle blunting, calcified aorta"*).

b. You will then generate a causal explanation using the following steps:

- Extract key diagnostic conclusions.
- Describe how the findings are interrelated.
- Retain the location of disorders, tubes, lines, and devices (if present), and provide the final causal reasoning.
- Present the most probable causes or implications for the patient's condition.

c. Please adhere to the sentence structure used by the radiologist.
Here are a few examples of Radiology Report and their corresponding Casual Explanation.