References

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, and Edward Choi. EHRXQA: A Multi-Modal Question Answering Dataset for Electronic Health Records with Chest X-ray Images. *Advances in Neural Information Processing Systems*, 36:3867–3880, December 2023. 2, 3.1, 5, C.3.3

Till J. Bungert, Levin Kobelke, and Paul F. Jäger. Understanding Silent Failures in Medical Image Classification. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 400–410, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43898-1. doi: 10.1007/978-3-031-43898-1_39. 3.1

Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17478-w. 2, 3.1, 3.1, 3.1, 3.1

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are We on the Right Way for Evaluating Large Vision-Language Models?, April 2024a. 2

Shuo Chen, Jindong Gu, Zhen Han, Yunpu Ma, Philip Torr, and Volker Tresp. Benchmarking robustness of adaptation methods on pre-trained vision-language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 51758–51777. Curran Associates, Inc., 2023. 1, 1, 2, 2, 2, 4.1

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis Langlotz. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation, January 2024b. 5

Cheng-Han Chiang and Hung-yi Lee. Can Large Language Models Be an Alternative to Human Evaluations? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.870. 2

Youngwon Choi, Wenxi Yu, Mahesh B. Nagarajan, Pangyu Teng, Jonathan G. Goldin, Steven S. Raman, Dieter R. Enzmann, Grace Hyun J. Kim, and Matthew S. Brown. Translating AI to Clinical Practice: Overcoming Data Shift with Explainability. *RadioGraphics*, 43(5):e220105, May 2023. ISSN 0271-5333. doi: 10.1148/rg.220105. 3.1

Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond Question-Based Biases: Assessing Multimodal Shortcut Learning in Visual Question Answering. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1554–1563, Montreal, QC, Canada, October 2021. IEEE. ISBN 978-1-66542-812-5. doi: 10.1109/ICCV48922.2021.00160. 2

Raman Dutt, Linus Ericsson, Pedro Sanchez, Sotirios A. Tsaftaris, and Timothy Hospedales. Parameter-Efficient Fine-Tuning for Medical Image Analysis: The Missed Opportunity, June 2024. 4.2, 5

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as You Desire. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6556–6576, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.365. 2

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. 2

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017. 2

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, October 2021. 4.1

Yefan Huang, Xiaoli Wang, Feiyan Liu, and Guofeng Huang. OVQA: A Clinically Generated Visual Question Answering Dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pp. 2924–2938, New York, NY, USA, July 2022. Association for Computing Machinery. ISBN 978-1-4503-8732-3. doi: 10.1145/3477495.3531724. 2, 3.1, C.3.2

Kristian Nørgaard Jensen and Barbara Plank. Fine-tuning vs From Scratch: Do Vision & Language Models Have Similar Capabilities on Out-of-Distribution Visual Question Answering? In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 1496–1508, Marseille, France, June 2022. European Language Resources Association. 2, 2

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, October 2023. 2

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, December 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0322-0. C.3.3

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x. C.3.3

Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20, October 2017. ISSN 1077-3142. doi: 10.1016/j.cviu.2017.06.005. 2

M. G. Kendall. The Treatment of Ties in Ranking Problems. *Biometrika*, 33(3):239–251, 1945. ISSN 0006-3444. doi: 10.2307/2332303. 3, 3.2

Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are Red, Violets are Blue... But Should VQA expect Them To? In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2775–2784, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-509-2. doi: 10.1109/CVPR46437.2021.00280. 2

Tom Kocmi and Christian Federmann. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz (eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 193–203, Tampere, Finland, June 2023. European Association for Machine Translation. 2

Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning, September 2021. 4.1

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28541–28564, December 2023. 1, 1, 2, 2, 4.1

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A Semantically-Labeled Knowledge-Enhanced Dataset For Medical Visual Question Answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654, April 2021a. doi: 10.1109/ISBI48211.2021.9434010. 2, 3.1, C.3.1

Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning Causal Semantic Representation for Out-of-Distribution Prediction. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6155–6170. Curran Associates, Inc., 2021b. 1

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 1950–1965. Curran Associates, Inc., 2022. 4.1

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, December 2023a. 2

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. 2

Jie Ma, Pinghui Wang, Dechen Kong, Zewei Wang, Jun Liu, Hongbin Pei, and Junzhou Zhao. Robust Visual Question Answering: Datasets, Methods, and Future Challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5575–5594, August 2024. ISSN 1939-3539. doi: 10.1109/TPAMI.2024.3366154. 5

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The Effect of Natural Distribution Shift on Question Answering Models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6905–6916. PMLR, November 2020. 2

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-Flamingo: A Multimodal Medical Few-shot Learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, pp. 353–367. PMLR, December 2023. 1, 2

Yang Nan, Huichi Zhou, Xiaodan Xing, and Guang Yang. Beyond the Hype: A dispassionate look at vision-language models in medical scenario, August 2024. 2, 2

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S. Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative Radiology Report Evaluation and Error Notation, May 2024. 2

Letitia Parcalabescu and Anette Frank. MM-SHAP: A Performance-agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4032–4059, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.223. 2

Jielin Qiu, Yi Zhu, Xingjian Shi, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. Benchmarking Robustness under Distribution Shift of Multimodal Image-Text Models. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, October 2022. 2, 2, 2

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, July 2021. 2

Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, James H. F. Rudd, Evis Sala, and Carola-Bibiane Schönlieb. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3):199–217, March 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00307-0. 1

Mélanie Roschewitz, Galvin Khara, Joe Yearsley, Nisha Sharma, Jonathan J. James, Éva Ambrózay, Adam Heroux, Peter Kecskemethy, Tobias Rijken, and Ben Glocker. Automatic correction of performance drift under acquisition shift in medical image classification. *Nature Communications*, 14(1):6608, October 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-42396-y. 3.1

Alexander Shirnin, Nikita Andreev, Sofia Potapova, and Ekaterina Artemova. Analyzing the Robustness of Vision & Language Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2751–2763, 2024. ISSN 2329-9304. doi: 10.1109/TASLP.2024.3399061. 2

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards Expert-Level Medical Question Answering with Large Language Models, May 2023. 1

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. VL-Adapter: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5227–5237, 2022. 2

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring Robustness to Natural Distribution Shifts in Image Classification. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18583–18599. Curran Associates, Inc., 2020. 1, 2

Zilong Wang, Xufang Luo, Xinyang Jiang, Dongsheng Li, and Lili Qiu. LLM-RadJudge: Achieving Radiologist-Level Evaluation for X-Ray Report Generation, April 2024. 2

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. PMC-LLaMA: Toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843, September 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocae045. 1

Joy T. Wu, Nkechinyere N. Agu, Ismini Lourentzou, Arjun Sharma, Joseph A. Paguio, Jasper S. Yao, Edward C. Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo A. Celi, and Mehdi Moradi. Chest ImaGenome Dataset for Clinical Reasoning, July 2021. C.3.3

Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A. Mazurowski, and Heung-Il Suk. Domain Generalization for Medical Image Analysis: A Survey, February 2024. 1, 5

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and LLMs evaluations. *Advances in Neural Information Processing Systems*, 36:58478–58507, 2023. 2

Hao Zhang, Wenqi Shao, Hong Liu, Yongqiang Ma, Ping Luo, Yu Qiao, and Kaipeng Zhang. AVIBench: Towards Evaluating the Robustness of Large Vision-Language Model on Adversarial Visual-Instructions, March 2024. 1, 2, 2

A Evaluation Details

A.1 Failures of traditional metrics

Examples of failures of traditional token-matching metrics are shown in Figure 7

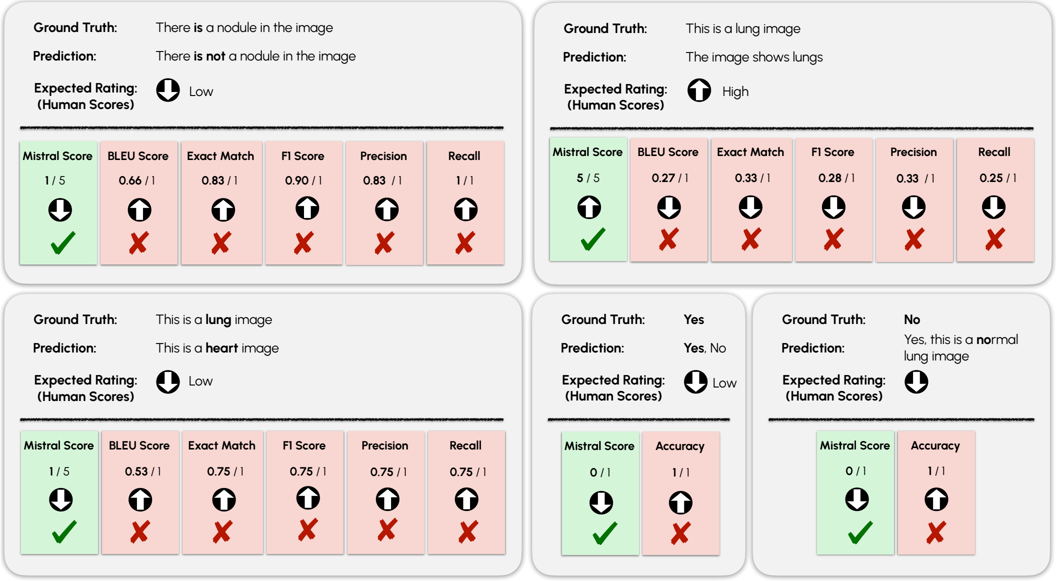

Ground Truth: There is a nodule in the image
Prediction: There is not a nodule in the image
Expected Rating: (Human Scores) Low
Mistral Score 1 / 5
BLEU Score 0.66 / 1
Exact Match 0.83 / 1
F1 Score 0.90 / 1
Precision 0.83 / 1
Recall 1 / 1
Ground Truth: This is a lung image
Prediction: The image shows lungs
Expected Rating: (Human Scores) High
Mistral Score 5 / 5
BLEU Score 0.27 / 1
Exact Match 0.33 / 1
F1 Score 0.28 / 1
Precision 0.33 / 1
Recall 0.25 / 1
Ground Truth: This is a lung image
Prediction: This is a heart image
Expected Rating: (Human Scores) Low
Mistral Score 1 / 5
BLEU Score 0.53 / 1
Exact Match 0.75 / 1
F1 Score 0.75 / 1
Precision 0.75 / 1
Recall 0.75 / 1
Ground Truth: Yes
Prediction: Yes, No
Expected Rating: (Human Scores) Low
Mistral Score 0 / 1
Accuracy 1 / 1
Ground Truth: No
Prediction: Yes, this is a normal lung image
Expected Rating: (Human Scores)
Mistral Score 0 / 1
Accuracy 1 / 1

Figure 7: Failures of Traditional Metrics

A.2 Prompts for Evaluation

Listing 1: Mistral Prompt for Evaluating Open-Ended Questions

```
<s>[INST] You are a helpful evaluator to evaluate answers to
    questions about biomedical images.
Score the following answer to a question about an image with
    respect to the ground truth answer with one to five stars.
Where the stars have the following meaning:
 1. One Star: "Incorrect"
   - The answer does not match the ground truth and contains
      significant inaccuracies.
   - Demonstrates a clear misunderstanding or misinterpretation
      of the question.
 2. Two Stars: "Partially Correct"
   - The answer has some elements that match the ground truth,
      but there are notable discrepancies.
   - Shows partial understanding but lacks overall accuracy in
      addressing the question.
 3. Three Stars: "Mostly Correct"
   - The answer aligns with the ground truth to a reasonable
      extent, but there are some inaccuracies or gaps.
   - Demonstrates a moderate understanding but may lack
 4. Four Stars: "Correct with Minor Deviations"
   - The answer is largely accurate and corresponds closely to
      the ground truth.
   - Minor deviations or omissions are present but do not
      significantly impact the overall correctness.
 5. Five Stars: "Perfect Match"
   - The answer exactly matches the ground truth with no
      discrepancies.
   - Demonstrates a precise and complete understanding of the
      question, providing a flawless response.
Here are some instructions on the input and output format:
 - The input will be passed as json format with the following
     fields that are important:
     - "question": the question about the image
     - "gt": the ground truth answer to the question
     - "pred": the predicted answer to the question
 - The output should be in json format and look the following:
    { mistralscore: <xxx>}
   where <xxx> is the number of stars you give to the answer.
       Do not add anything else to the answer.
 [/INST]
</s>
```

Listing 2: Mistral Prompt for Evaluating Closed-Ended Questions

```
<s>[INST] You are a helpful evaluator to evaluate answers to
    questions about biomedical images.
Score the following answer to a question about an image with
    respect to the ground truth answer with zero or one star.
The questions are all close ended, therefore the answer is
    either correct or false, there are no states in between.
Where the stars have the following meaning:
 0. Zero Star: "Incorrect"
   - The answer does not match the ground truth and contains
       significant inaccuracies.
   - Demonstrates a clear misunderstanding or misinterpretation
        of the question.
 1. One Star: "Perfect Match"
   - The answer exactly matches the ground truth with no
       discrepancies.
   - Demonstrates a precise and complete understanding of the
       question, providing a flawless response.
Here are some instructions on the input and output format:
 - The input will be passed as json format with the following
     fields that are important:
     - "question": the question about the image
     - "gt": the ground truth answer to the question
     - "pred": the predicted answer to the question
 - The output should be in json format and look the following:
    { mistralscore: <xxx>}
   where <xxx> is the number of stars you give to the answer.
       Do not add anything else to the answer.
 [/INST]
</s>
```

Listing 3: Mistral Prompt for Evaluating Closed-Ended Multilabel Questions

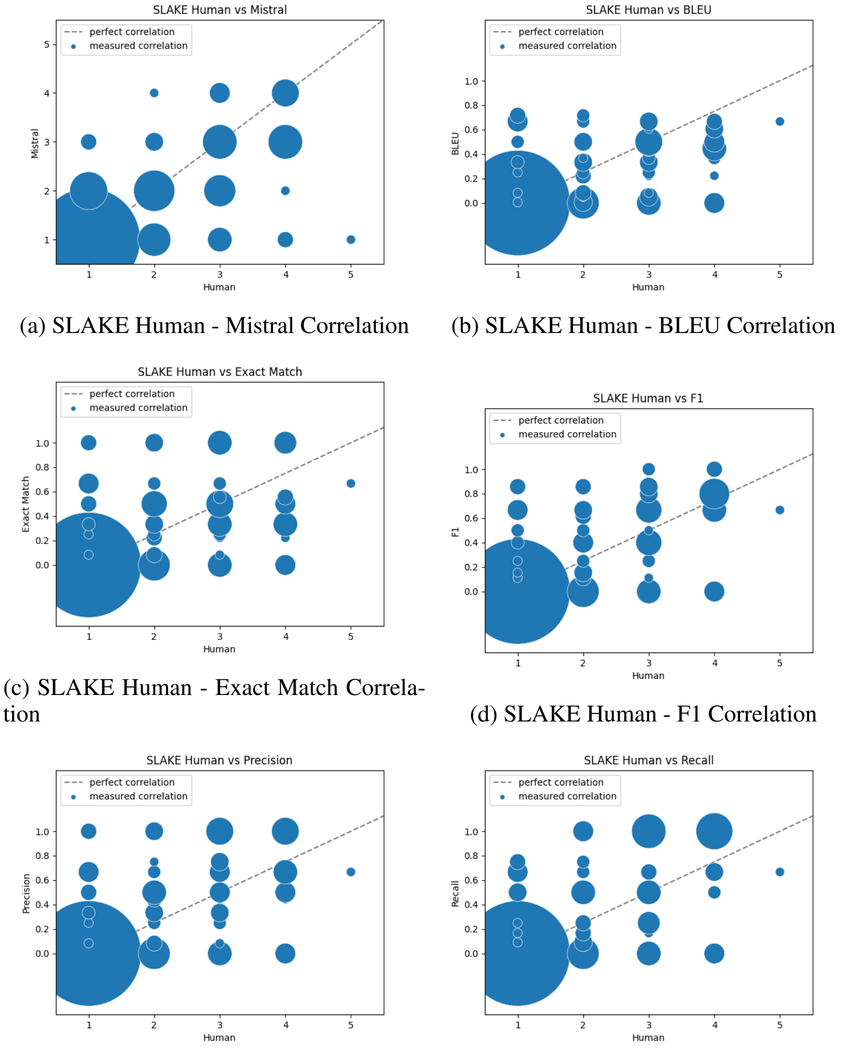
```
<s>[INST] You are a helpful evaluator to evaluate answers to
    questions about biomedical images.
Score the following answer to a question about an image with
    respect to the ground truth answer with 0, 0.5 or 1 star.
Each question asks for two options in the image and the answer
    can either be one of the options, both of the options or
    none.
The stars for rating have the following meaning:
 0 Star: "Incorrect"
   - The answer does not match the ground truth and contains
       significant inaccuracies.
   - Demonstrates a clear misunderstanding or misinterpretation
        of the question.
   - This is the case if
     - Option A is the ground truth answer, but the prediction
         is Option B
     - Option B is the ground truth answer, but the prediction
         is Option A
     - The ground truth answer is "both", but the prediction is
          "none"
     - The ground truth answer is "none", but the prediction is
          "both"
 0.5 Star: "Partially Correct"
```

```
   - The answer partially matches the ground truth, but
       contains some inaccuracies.
   - Demonstrates a partial understanding of the question,
       providing a partially correct response.
   - This is the case if
     - Option A/B is the ground truth answer, but the
         prediction is "both"
     - Option A/B is the ground truth answer, but the
         prediction is "none"
     - The ground truth is "both", but the prediction is option
          A/B
     - The ground truth in "none", but the prediction is option
          A/B
 1 Star: "Perfect Match"
   - The answer exactly matches the ground truth with no
       discrepancies.
   - Demonstrates a precise and complete understanding of the
       question, providing a flawless response.
   - This is the case if
     - Option A is the ground truth answer and the prediction
         is Option A
     - Option B is the ground truth answer and the prediction
         is Option B
     - The ground truth is "both" and the prediction is "both"
     - The ground truth is "none" and the prediction is "none"

Especially for the "none" Cases:
    When the ground truth is "none":
        If the prediction is "none", the score should be 1 star
            .
        If the prediction is "both", the score should be 0
            stars.
        If the prediction is Option A or B, the score should be
             0.5 stars.
    When the prediction is "none":
        If the ground truth is "none", the score should be 1
            star.
        If the ground truth is "both", the score should be 0
            stars.
        If the ground truth is Option A or B, the score should
            be 0.5 stars.

Especially for the "both" Cases:
    When the ground truth is "both":
        If the prediction is "both", the score should be 1 star
            .
        If the prediction is "none", the score should be 0
            stars.
        If the prediction is Option A or B, the score should be
             0.5 stars.
    When the prediction is "both":
        If the ground truth is "both", the score should be 1
            star.
        If the ground truth is "none", the score should be 0
            stars.
        If the ground truth is Option A or B, the score should
            be 0.5 stars.

Here are some instructions on the input and output format:
 - The input will be passed as json format with the following
     fields that are important:
    - "question": the question about the image
    - "gt": the ground truth answer to the question
    - "pred": the predicted answer to the question
 - The output should be in json format and look the following:
    { mistralscore: <xxx>}
   where <xxx> is the number of stars you give to the answer.
      Do not add anything else to the answer.
 [/INST]
</s>
```

B Human Rater Study Details

B.1 Detailed Results of the Human Rater Study

The following figures show detailed results of the human rater study. Figure 8, Figure 9, and Figure 10 show scatter plots with the correlation between the human ratings and the other metrics. Figure 11 shows detailed correlation results, including the correlation between Mistral and the other metrics.



(a) SLAKE Human - Mistral Correlation

(b) SLAKE Human - BLEU Correlation

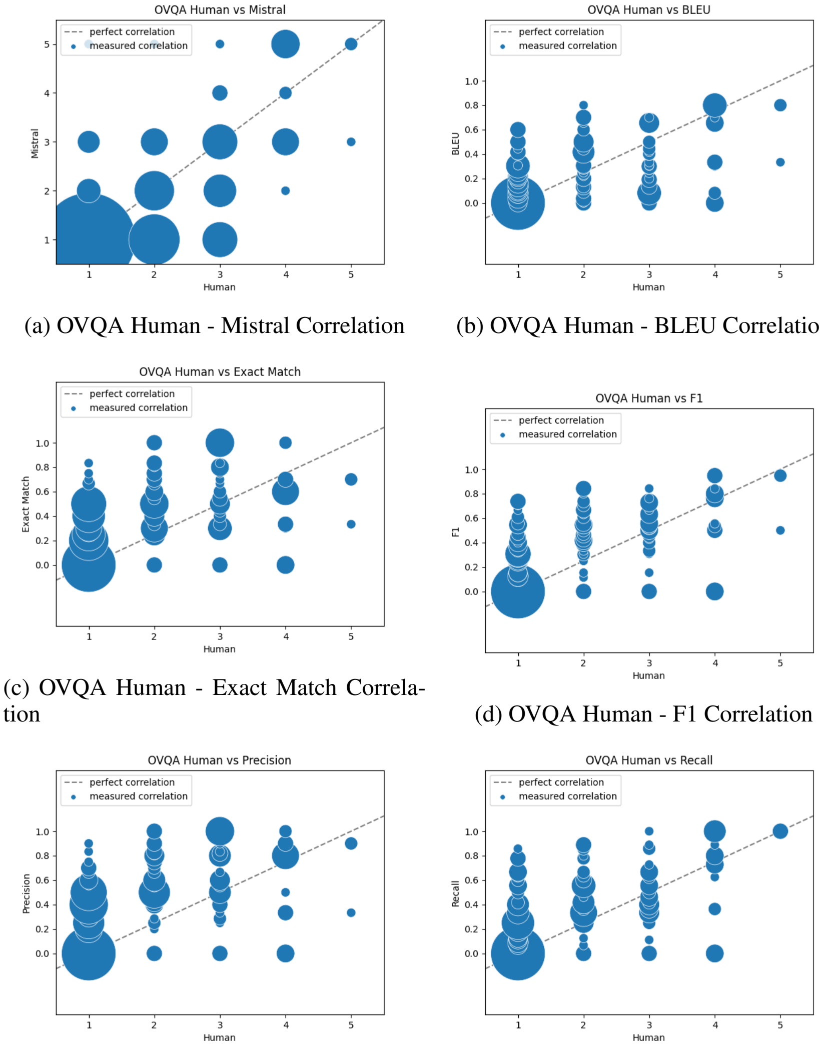
(c) SLAKE Human - Exact Match Correlation

(d) SLAKE Human - F1 Correlation

(e) SLAKE Human - Precision Correlation

(f) SLAKE Human - Recall Correlation

Figure 8: Scatter plots showing the correlation between the human ratings and respective other metrics on the SLAKE dataset. Size of the dots indicates the number of ratings that correspond to that point.



(a) OVQA Human - Mistral Correlation

(b) OVQA Human - BLEU Correlation

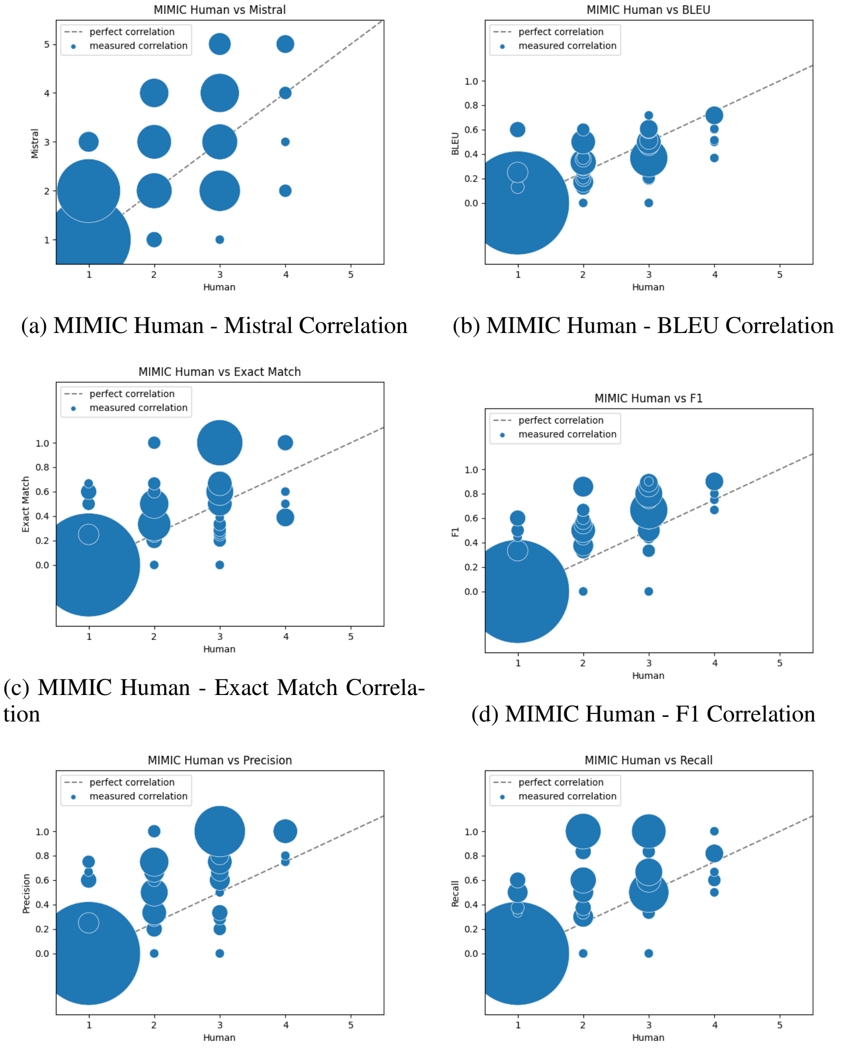
(c) OVQA Human - Exact Match Correlation

(d) OVQA Human - F1 Correlation

(e) OVQA Human - Precision Correlation

(f) OVQA Human - Recall Correlation

Figure 9: Scatter plots showing the correlation between the human ratings and respective other metrics on the OVQA dataset. Size of the dots indicates the number of ratings that correspond to that point.



(a) MIMIC Human - Mistral Correlation

(b) MIMIC Human - BLEU Correlation

(c) MIMIC Human - Exact Match Correlation

(d) MIMIC Human - F1 Correlation

(e) MIMIC Human - Precision Correlation

(f) MIMIC Human - Recall Correlation

Figure 10: Scatter plots showing the correlation between the human ratings and respective other metrics on the MIMIC dataset. Size of the dots indicates the number of ratings that correspond to that point.

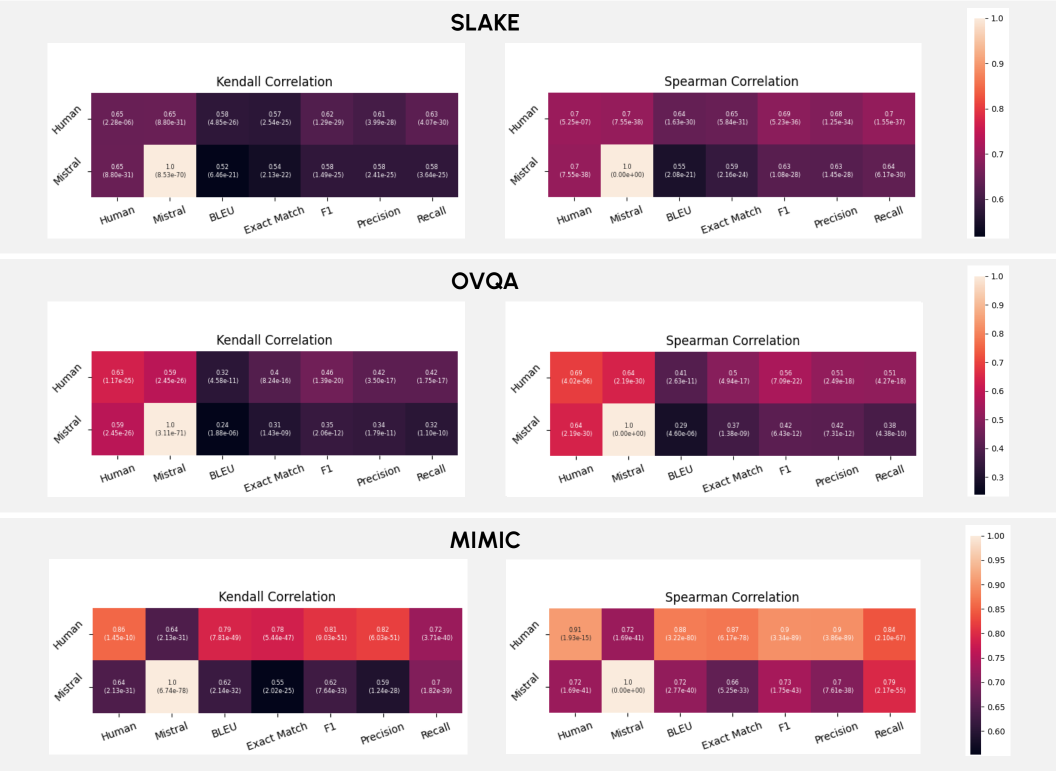


Figure 11: **Extended Results of the Human Rater Study**. Human interrater correlation is calculated between two human raters. Shown are the Kendall and Spearman correlation between the human rating and all traditional metrics as well as the correlation between Mistral and the traditional metrics.

C Robustness Study Details

C.1 Hyperparameter Search

We performed several hyperparameter sweeps for each dataset and PEFT method in order to find suitable setups for the experiments in the PEFT robustness study. For the hyperparameter sweeps, we trained on the whole training set for each dataset and PEFT method and ran inference on the validation set. Training ran for 3 epochs and 3 seeds for each experiment.

C.1.1 Prompt Tuning

For prompt tuning, we performed the following hyperparameter sweeps:

- Number of tokens: $[40, 60, 80, 100]$
- Learning rate: $[3e-2, 3e-1]$

The results for SLAKE can be found in Table 1, for OVQA in Table 2, and for MIMIC in Table 3.

Table 1: Hyperparameter sweep for prompt tuning on the SLAKE dataset. Selected hyperparameters for the final PEFT robustness study are highlighted. Mean and standard deviation are reported for three seeds.

# Tokens	Learning Rate	Closed-Ended (Mistral Accuracy)	Open-Ended (Mistral Score)
40	3e-2	0.79 +/- 0.02	4.17 +/- 0.04
40	3e-1	0.8 +/- 0.02	4.19 +/- 0.05
60	3e-2	0.76 +/- 0.04	4.17 +/- 0.03
60	3e-1	0.81 +/- 0.01	4.17 +/- 0.05
80	3e-2	0.78 +/- 0.05	4.15 +/- 0.01
80	3e-1	0.82 +/- 0.01	4.18 +/- 0.02
100	3e-2	0.77 +/- 0.06	4.15 +/- 0.05
100	3e-1	0.81 +/- 0.0	4.18 +/- 0.03

Table 2: Hyperparameter sweep for prompt tuning on the OVQA dataset. Selected hyperparameters for the final PEFT robustness study are highlighted. Mean and standard deviation are reported for three seeds.

# Tokens	Learning Rate	Closed-Ended (Mistral Accuracy)	Open-Ended (Mistral Score)
40	3e-2	0.85 +/- 0.0	2.96 +/- 0.06
40	3e-1	0.85 +/- 0.01	2.95 +/- 0.05
60	3e-2	0.85 +/- 0.01	2.98 +/- 0.04
60	3e-1	0.85 +/- 0.0	2.97 +/- 0.04
80	3e-2	0.81 +/- 0.04	2.96 +/- 0.04
80	3e-1	0.84 +/- 0.01	2.99 +/- 0.02
100	3e-2	0.83 +/- 0.02	2.99 +/- 0.04
100	3e-1	0.85 +/- 0.0	3.0 +/- 0.03

Table 3: Hyperparameter sweep for prompt tuning on the MIMIC dataset. Selected hyperparameters for the final PEFT robustness study are highlighted. Mean and standard deviation are reported for three seeds.

# Tokens	Learning Rate	Closed-Ended (Mistral Accuracy)	Open-Ended (Mistral Score)
40	3e-2	0.67 +/- 0.02	3.17 +/- 0.03
40	3e-1	0.67 +/- 0.01	3.15 +/- 0.04
60	3e-2	0.68 +/- 0.01	3.14 +/- 0.01
60	3e-1	0.69 +/- 0.01	3.19 +/- 0.02
80	3e-2	0.66 +/- 0.05	3.17 +/- 0.01
80	3e-1	0.68 +/- 0.02	3.17 +/- 0.03
100	3e-2	0.68 +/- 0.02	3.19 +/- 0.03
100	3e-1	0.67 +/- 0.01	3.17 +/- 0.03

C.1.2 LoRA

For LoRA, we performed the following hyperparameter sweeps:

- Rank: $[16, 32, 64, 128, 256]$
- Learning rate: $[3e-5, 3e-4]$

α is set to 2 × Rank. The results for SLAKE can be found in Table 4, for OVQA in Table 5, and for MIMIC in Table 6. Note that some of the hyperparameter configurations led to instabilities during training loss, indicated by ”NaN”.

Table 4: Hyperparameter sweep for LoRA on the SLAKE dataset. Selected hyperparameters for the final PEFT robustness study are highlighted. Mean and standard deviation are reported for three seeds. Rows with ”NaN” showed instabilities in the loss during training.

Rank	Learning Rate	Closed-Ended (Mistral Accuracy)	Open-Ended (Mistral Score)
16	3e-5	0.83 +/- 0.01	4.24 +/- 0.02
16	3e-4	0.82 +/- 0.01	4.23 +/- 0.05
32	3e-5	0.85 +/- 0.01	4.27 +/- 0.04
32	3e-4	0.73 +/- 0.07	4.2 +/- 0.03
64	3e-5	0.84 +/- 0.01	4.29 +/- 0.04
64	3e-4	0.52 +/- 0.07	3.01 +/- 1.28
128	3e-5	0.84 +/- 0.01	4.31 +/- 0.03
128	3e-4	NaN	NaN
256	3e-5	0.83 +/- 0.01	4.28 +/- 0.01
256	3e-4	NaN	NaN

Table 5: Hyperparameter sweep for LoRA on the OVQA dataset. Selected hyperparameters for the final PEFT robustness study are highlighted. Mean and standard deviation are reported for three seeds. Rows with ”NaN” showed instabilities in the loss during training.

Rank	Learning Rate	Closed-Ended (Mistral Accuracy)	Open-Ended (Mistral Score)
16	3e-5	0.84 +/- 0.0	3.02 +/- 0.07
16	3e-4	0.83 +/- 0.02	3.08 +/- 0.03
32	3e-5	0.85 +/- 0.0	3.04 +/- 0.01
32	3e-4	0.82 +/- 0.01	2.99 +/- 0.04
64	3e-5	0.85 +/- 0.0	3.11 +/- 0.02
64	3e-4	0.65 +/- 0.0	2.04 +/- 0.1
128	3e-5	0.85 +/- 0.0	3.09 +/- 0.04
128	3e-4	NaN	NaN
256	3e-5	0.85 +/- 0.0	3.1 +/- 0.03
256	3e-4	NaN	NaN

Table 6: Hyperparameter sweep for LoRA on the MIMIC dataset. Selected hyperparameters for the final PEFT robustness study are highlighted. Mean and standard deviation are reported for three seeds. Rows with ”NaN” showed instabilities in the loss during training.

Rank	Learning Rate	Closed-Ended (Mistral Accuracy)	Open-Ended (Mistral Score)
16	3e-5	0.7 +/- 0.01	3.31 +/- 0.01
16	3e-4	0.68 +/- 0.01	3.18 +/- 0.04
32	3e-5	0.71 +/- 0.0	3.33 +/- 0.02
32	3e-4	0.42 +/- 0.16	2.34 +/- 0.06
64	3e-5	0.71 +/- 0.01	3.33 +/- 0.03
64	3e-4	NaN	NaN
128	3e-5	0.7 +/- 0.0	3.35 +/- 0.04
128	3e-4	NaN	NaN
256	3e-5	NaN	NaN
256	3e-4	NaN	NaN

C.2 $(\texttt{IA})^3$

For $(\texttt{IA})^3$, we performed the following hyperparameter sweeps:

- Learning rate: $[3e - 3, 3e - 2, 3e - 1]$

The results for SLAKE can be found in Table 7, for OVQA in Table 8, and for MIMIC in Table 9.

Table 7: Hyperparameter sweep for $(\texttt{IA})^3$on the SLAKE dataset. Selected hyperparameters for the final PEFT robustness study are highlighted. Mean and standard deviation are reported for three seeds.

Learning Rate	Closed-Ended (Mistral Accuracy)	Open-Ended (Mistral Score)
lr3e-3	0.63 +/- 0.02	3.74 +/- 0.02
lr3e-2	0.83 +/- 0.01	4.28 +/- 0.02
lr3e-1	0.65 +/- 0.01	4.21 +/- 0.05

Table 8: Hyperparameter sweep for $(\texttt{IA})^3$on the OVQA dataset. Selected hyperparameters for the final PEFT robustness study are highlighted. Mean and standard deviation are reported for three seeds.

Learning Rate	Closed-Ended (Mistral Accuracy)	Open-Ended (Mistral Score)
lr3e-3	0.75 +/- 0.01	2.84 +/- 0.02
lr3e-2	0.84 +/- 0.0	3.08 +/- 0.01
lr3e-1	0.78 +/- 0.04	2.97 +/- 0.05

Table 9: Hyperparameter sweep for $(\texttt{IA})^3$on the MIMIC dataset. Selected hyperparameters for the final PEFT robustness study are highlighted. Mean and standard deviation are reported for three seeds.

Learning Rate	Closed-Ended (Mistral Accuracy)	Open-Ended (Mistral Score)
lr3e-3	0.53 +/- 0.0	2.86 +/- 0.01
lr3e-2	0.7 +/- 0.01	3.3 +/- 0.04
lr3e-1	0.61 +/- 0.05	3.06 +/- 0.04

C.3 Dataset Details

C.3.1 SLAKE

The SLAKE dataset Liu et al. (2021a) is a bilingual radiological VQA dataset, containing English and Chinese questions. We use the English subset of the SLAKE dataset. The dataset is composed of MRI, CT, and X-ray images. All images are 2D, so for the MRI and CT images, single slices are extracted. For each question, metadata information about the location, the modality, and the content is provided. Overall, the images are split into 5 different body locations, 11 different content types (question types), and the mentioned three modalities.

The exact sizes of the dataset splits are listed in Table 10. Note that for the modality shift, we merged the test set with the OoD cases from the training set, since the images are distinct, and thus, the same image cannot appear in the training and test set. As this is not the case for the question type shift, we only use the OoD cases from the test set here.

Table 10: Size of the SLAKE dataset for the different splits.

Split	i.i.d./OoD/all	# Cases
Whole Dataset		
Train	all	4866
Validate	all	1043
Modality Shift (OoD: X-Ray)		
Train	i.i.d.	3448
Test	i.i.d.	689
Test	OoD	1779
Question Type Shift (OoD: Size)		
Train	i.i.d.	4581
Test	i.i.d.	994
Test	OoD	56

C.3.2 OVQA

The OVQA dataset Huang et al. (2022) is an orthopedic VQA dataset, containing CT and X-Ray images. All images are 2D, so for the CT images, either a 3D rendering is shown as a 2D image or a single plane. For each question, metadata information is provided about the imaged organ (like the ”location” in the SLAKE dataset), and the question type (like the ”content” in SLAKE) is provided. The dataset contains 6 different question types and 4 different body parts.

The exact sizes of the dataset splits are listed in Table 11. We removed closed-ended questions with more than two categories to choose from and closed-ended questions where the categories to answer were not exactly contained in the question. As for the SLAKE dataset, we merged the questions from the training set to the OoD test set for the organ shift, but not for the question type shift.

Table 11: Size of the OVQA dataset for the different splits.

Split	i.i.d./OoD/all	# Cases
Whole Dataset		
Train	all	13492
Validate	all	1645
Organ Shift (OoD: Leg)		
Train	i.i.d.	8755
Test	i.i.d.	1044
Test	OoD	5350
Question Type Shift (OoD: Organ System)		
Train	i.i.d.	11924
Test	i.i.d.	1420
Test	OoD	237

C.3.3 MIMIC-CXR-VQA

The MIMIC-CXR-VQA dataset Bae et al. (2023) is a chest X-ray dataset, which is built based on the MIMIC-CXR dataset Johnson et al. (2019), the MIMIC-IV dataset Johnson et al. (2023), and the Chest ImaGenome dataset Wu et al. (2021). For each question, the semantic type is specified. Three different semantic types are specified, which are ”choose”, ”query”, and ”verify”. For ”choose”, the

task is to choose between two options provided in the answer, but also both or none of the options can be correct. For ”query”, the task is to list all the categories that match the questions, e.g. all anatomical findings. Lastly, ”verify” are yes/no questions. All the questions can be answered based on a fixed set of classes, where the dataset overall contains 110 answer labels. The answers are given as a list of the correct classes. We preprocess the questions differently, based on their semantic type: For the ”choose” questions, whenever the list of answers contains both options, we change the answer to ”both”, and whenever the list of answers is empty, we change the answer to ”none”. For the ”query” questions, we concatenate the list of answers to one string, with the answer labels being comma-separated. For the ”verify” questions, we do not apply any specific preprocessing.
The information for the patient’s gender, ethnicity, and age are taken from the MIMIC-IV dataset. Whenever the metadata information of a subject ID is not unique, we set it to ”none”. In the respective shifts, we exclude questions where the corresponding metadata field is not known, which includes all fields with ”none”, and for the ethnicity shift also the value ”unknown/other”. The exact sizes of the dataset splits are listed in Table 12.

Table 12: Size of the MIMIC dataset for the different splits.

Split	i.i.d./OoD/all	# Cases
Whole Dataset		
Train	all	290031
Validate	all	73567
Gender Shift (OoD: Female)		
Train	i.i.d.	147790
Test	i.i.d.	7277
Test	OoD	6120
Ethnicity Shift (OoD: Non-white)		
Train	i.i.d.	171593
Test	i.i.d.	8101
Test	OoD	3713
Age Shift (OoD: Young)		
Train	i.i.d.	155941
Test	i.i.d.	6686
Test	OoD	2076

C.3.4 Ratio of unique questions in the datasets

Table 13: Ratio of unique questions in the datasets

		Overall	Unique	Ratio
Train	**MIMIC**	290031	132387	0.46
	SLAKE	4866	579	0.12
	OVQA	13492	960	0.07
Val	**MIMIC**	73567	31148	0.42
	SLAKE	1043	314	0.3
	OVQA	1645	266	0.16
Test	**MIMIC**	13793	7565	0.55
	SLAKE	1050	313	0.3
	OVQA	1657	335	0.2

C.4 Detailed Results of the Robustness Study

Tables 14-19 show the detailed results of the robustness study. Further, Figure 12 shows the inter-method and inter-shift variability of the different PEFT methods, so not including full FT.

Table 14: **Robustness Results on the SLAKE Dataset.** Results with ± indicate the mean and standard deviation over three seeds. Note that the most frequent baseline can only be calculated for the i.i.d. set as for OoD too few questions match the training set. RR: Relative Robustness.

	Modality Shift OoD: X-Ray						Question Type Shift OoD: Size					
	Closed-Ended			Open-Ended			Closed-Ended			Open-Ended		
	i.i.d.	OoD	RR	i.i.d.	OoD	RR	i.i.d.	OoD	RR	i.i.d.	OoD	RR
No Finetune	0.59	0.29	0.49	2.91	2.79	0.96	0.54	0.47	0.87	2.86	3.16	1.11
Full Finetune	0.57	0.48	0.83	4.03	3.17	0.79	0.56	0.35	0.63	4.11	4.38	1.07
Prompt	0.85±0.01	0.62±0.03	0.73±0.05	4.22±0.04	3.69±0.06	0.87±0.02	0.85±0.01	0.49±0.12	0.57±0.14	4.17±0.01	4.35±0.16	1.04±0.04
LoRA	0.88±0.0	0.45±0.04	0.51±0.04	4.34±0.06	3.61±0.06	0.83±0.03	0.87±0.0	0.39±0.07	0.45±0.08	4.26±0.01	4.26±0.07	1.0±0.02
$(\texttt{IA})^3$	0.85±0.01	0.64±0.07	0.75±0.08	4.35±0.02	3.4±0.15	0.78±0.04	0.87±0.02	0.53±0.06	0.61±0.06	4.23±0.01	4.21±0.27	0.99±0.07
Most Freq.	0.69	-	-	3.22	-	-	0.696	-	-	3.05	-	-

Table 15: **No Image Baseline on the SLAKE Dataset.** Results with ± indicate the mean and standard deviation over three seeds. The model was trained with the same methods as Table 14 just without seeing the image content. RR: Relative Robustness.

	Modality Shift OoD: X-Ray						Question Type Shift OoD: Size					
	Closed-Ended			Open-Ended			Closed-Ended			Open-Ended		
	i.i.d.	OoD	RR	i.i.d.	OoD	RR	i.i.d.	OoD	RR	i.i.d.	OoD	RR
No Finetune	0.46	0.35	0.75	2.13	2.33	1.1	0.46	0.29	0.64	2.22	2.03	0.91
Prompt	0.55±0.01	0.5±0.01	0.91±0.03	3.24±0.0	1.88±0.06	0.58±0.02	0.59±0.05	0.55±0.07	0.94±0.18	3.18±0.04	2.26±0.6	0.71±0.19
LoRA	0.64±0.03	0.47±0.07	0.73±0.11	3.24±0.02	1.93±0.05	0.6±0.02	0.69±0.01	0.53±0.18	0.76±0.25	3.12±0.03	2.95±0.0	0.94±0.01
$(\texttt{IA})^3$	0.55±0.01	0.47±0.02	0.85±0.03	3.26±0.01	1.87±0.02	0.57±0.01	0.55±0.08	0.53±0.06	0.96±0.09	3.15±0.02	2.64±0.53	0.84±0.17

Table 16: **Robustness Results on the OVQA Dataset.** Results with ± indicate the mean and standard deviation over three seeds. Note that the most frequent baseline can only be calculated for the i.i.d. set as for OoD too few questions match the training set. RR: Relative Robustness.

	Body Part Shift OoD: Leg						Question Type Shift OoD: Organ System					
	Closed-Ended			Open-Ended			Closed-Ended			Open-Ended		
	i.i.d.	OoD	RR	i.i.d.	OoD	RR	i.i.d.	OoD	RR	i.i.d.	OoD	RR
No Finetune	0.42	0.4	0.96	2.4	2.45	1.02	0.43	0.33	0.75	2.39	1.94	0.81
Full Finetune	0.7	0.55	0.77	3.16	2.23	0.71	0.76	0.08	0.1	2.97	1.02	0.34
Prompt	0.86±0.0	0.75±0.01	0.87±0.01	3.12±0.02	2.38±0.02	0.76±0.01	0.82±0.04	0.86±0.01	1.05±0.06	2.9±0.01	1.7±0.02	0.59±0.01
LoRA	0.86±0.01	0.77±0.0	0.9±0.01	3.23±0.02	2.47±0.05	0.76±0.02	0.84±0.0	0.74±0.06	0.88±0.07	3.09±0.03	1.72±0.11	0.56±0.04
$(\texttt{IA})^3$	0.83±0.02	0.75±0.01	0.9±0.02	3.21±0.05	2.46±0.04	0.77±0.0	0.8±0.02	0.72±0.11	0.91±0.15	2.98±0.06	1.51±0.05	0.51±0.02
Most Freq.	0.75	-	-	2.57	-	-	0.73	-	-	2.23	-	-

Table 17: **No Image Baseline on the OVQA Dataset.** Results with ± indicate the mean and standard deviation over three seeds. The model was trained with the same methods as Table 16 just without seeing the image content. RR: Relative Robustness.

	Body Part Shift OoD: Leg						Question Type Shift OoD: Organ System					
	Closed-Ended			Open-Ended			Closed-Ended			Open-Ended		
	i.i.d.	OoD	RR	i.i.d.	OoD	RR	i.i.d.	OoD	RR	i.i.d.	OoD	RR
No Finetune	0.42	0.36	0.85	1.37	2.02	1.47	0.41	0.4	0.97	1.39	1.29	0.93
Prompt	0.74±0.01	0.69±0.02	0.93±0.01	2.63±0.1	2.12±0.07	0.81±0.01	0.67±0.03	0.44±0.01	0.66±0.02	2.35±0.06	1.26±0.07	0.54±0.02
LoRA	0.74±0.0	0.7±0.0	0.95±0.01	2.67±0.16	2.14±0.01	0.81±0.05	0.73±0.0	0.43±0.03	0.6±0.04	2.36±0.02	1.29±0.09	0.55±0.04
$(\texttt{IA})^3$	0.74±0.0	0.69±0.02	0.93±0.02	2.74±0.07	2.13±0.04	0.78±0.03	0.7±0.04	0.39±0.06	0.56±0.1	2.36±0.02	1.28±0.06	0.54±0.02

Table 18: **Robustness Results on the MIMIC Dataset.** Results with ± indicate the mean and standard deviation over three seeds. Note that the most frequent baseline can not be calculated as too few questions match the training set. RR: Relative Robustness.

	Gender Shift OoD: Female						Ethnicity Shift OoD: Non-white						Age Shift OoD: Young					
	Closed-Ended			Open-Ended			Closed-Ended			Open-Ended			Closed-Ended			Open-Ended		
	i.i.d.	OoD	RR	i.i.d.	OoD	RR	i.i.d.	OoD	RR	i.i.d.	OoD	RR	i.i.d.	OoD	RR	i.i.d.	OoD	RR
No Finetune	0.51	0.49	0.97	2.36	2.31	0.98	0.5	0.49	0.98	2.37	2.34	0.99	0.51	0.47	0.92	2.36	2.36	1
Full Finetune	0.75	0.75	1.01	3.41	3.44	1.01	0.72	0.77	1.07	3.27	3.51	1.07	0.71	0.79	1.12	3.3	3.45	1.05
Prompt	0.52±0.06	0.54±0.05	1.04±0.03	3.25±0.05	3.3±0.05	1.01±0.0	0.54±0.14	0.64±0.14	1.19±0.05	3.14±0.11	3.37±0.28	1.07±0.05	0.51±0.14	0.66±0.07	1.37±0.35	3.19±0.03	3.35±0.08	1.05±0.01
LoRA	0.75±0.01	0.76±0.0	1.02±0.01	3.35±0.11	3.41±0.06	1.02±0.01	0.71±0.03	0.79±0.02	1.12±0.02	3.32±0.04	3.58±0.01	1.08±0.01	0.73±0.01	0.78±0.01	1.07±0.0	3.36±0.04	3.54±0.1	1.05±0.02
$(\texttt{IA})^3$	0.63±0.1	0.65±0.08	1.04±0.04	3.32±0.04	3.36±0.05	1.01±0.01	0.6±0.06	0.7±0.05	1.16±0.02	3.26±0.04	3.51±0.0	1.08±0.02	0.52±0.02	0.72±0.06	1.39±0.05	3.18±0.09	3.34±0.21	1.05±0.04

Table 19: **No Image Baseline on the MIMIC Dataset.** Results with ± indicate the mean and standard deviation over three seeds. The model was trained with the same methods as Table 18 just without seeing the image content. RR: Relative Robustness.

	Gender Shift OoD: Female						Ethnicity Shift OoD: Non-white						Age Shift OoD: Young					
	Closed-Ended			Open-Ended			Closed-Ended			Open-Ended			Closed-Ended			Open-Ended		
	i.i.d.	OoD	RR	i.i.d.	OoD	RR	i.i.d.	OoD	RR	i.i.d.	OoD	RR	i.i.d.	OoD	RR	i.i.d.	OoD	RR
No Finetune	0.49	0.48	0.97	2.34	2.37	1.01	0.49	0.48	0.97	2.37	2.39	1.01	0.51	0.46	0.89	2.33	2.49	1.07
Prompt	0.53±0.0	0.5±0.0	0.95±0.0	2.71±0.01	2.63±0.02	0.97±0.01	0.54±0.0	0.42±0.0	0.78±0.0	2.71±0.04	2.43±0.02	0.9±0.01	0.6±0.0	0.34±0.01	0.57±0.01	2.89±0.03	2.3±0.02	0.8±0.0
LoRA	0.53±0.0	0.5±0.0	0.95±0.0	2.74±0.01	2.64±0.01	0.97±0.0	0.54±0.0	0.41±0.0	0.76±0.0	2.75±0.02	2.43±0.02	0.88±0.01	0.59±0.01	0.35±0.02	0.6±0.04	2.93±0.04	2.25±0.02	0.77±0.01
$(\texttt{IA})^3$	0.53±0.01	0.51±0.01	0.95±0.01	2.72±0.03	2.64±0.02	0.97±0.01	0.54±0.0	0.42±0.0	0.77±0.0	2.71±0.02	2.46±0.02	0.91±0.01	0.6±0.0	0.34±0.0	0.57±0.01	2.91±0.02	2.29±0.04	0.79±0.01

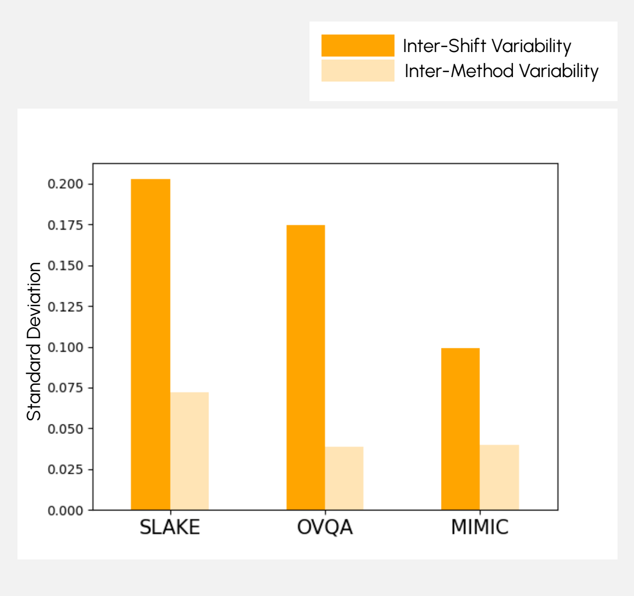


Figure 12: Standard deviation between shifts vs. standard deviation between PEFT methods, not including full FT. The type of shift has a higher impact on the robustness than the PEFT method.

D Corruption Study

We compared the realistic shifts defined for our datasets (R1) with artificial shifts, meaning image corruptions, to assess whether artificial shifts correspond to real-world shifts. The artificial data shifts were generated through image corruptions, including blur, Gaussian noise, and brightness adjustments. They were applied in different strengths (low, medium, and high). We used OpenCV for the image corruptions with the settings shown in Table 20.

Table 20: Corruption settings for the artificial shifts. Brackets indicate the altered parameter for each corruption, [...,...] indicate ranges for the corruption where randomly a value in that range is chosen.

	Blur (Kernel Size)	Gaussian Noise (Mean)	Brightness (Alpha)
Low	5	[0, 0.06]	[1.1, 2]
Medium	7	[0.09, 0.15]	[2.5, 4]
High	11	[0.18, 0.25]	[4.5, 6]

For this sample study, we used the LLaVA-Med model fine-tuned on the SLAKE dataset with the $(\texttt{IA})^3$method. The i.i.d. and OoD samples for realistic shifts were as previously described (R1). For artificial shifts, the i.i.d. train and test samples were identical to those used for realistic shifts, while OoD test samples were created by corrupting the i.i.d. test images with varying strengths of blur, brightness, and noise. Each corruption method was applied with a probability of 0.5, with at least one corruption always being applied.

Table 21 shows the relative robustness results for both artificial and realistic shifts. The results show that both modality shift and question type shift exhibit lower relative robustness compared to all artificial shifts at low, medium, and high strengths. This suggests that artificial shifts, such as image corruption, fail to accurately represent the challenges posed by real-world, realistic shifts. The most prominent example here is the relative robustness of closed-ended questions under the question type shift (realistic shift), which is up to 96% compared to the realistic shift which only has 61%. The only exception where the realistic shift shows higher robustness is the question type shift on the open-ended questions, which is already nearly 100% on the realistic shift.

Table 21: Robustness results for the artificial and realistic shifts on SLAKE dataset

	Corruption Shift (OoD: Corrupted i.i.d images)						Corruption Shift (OoD: Corrupted i.i.d images)					
	Closed Ended			Open Ended			Closed Ended			Open Ended		
	i.i.d.	**OoD**	**RR**	**i.i.d.**	**OoD**	**RR**	**i.i.d.**	**OoD**	**RR**	**i.i.d.**	**OoD**	**RR**
Low Corruption	0.85±0.01	0.83±0.01	0.98±0.0	4.35±0.02	4.21±0.05	0.97±0.02	0.87±0.02	0.84±0.0	0.96±0.02	4.23±0.01	4.16±0.03	0.98±0.01
Medium Corruption	0.85±0.01	0.79±0.02	0.94±0.02	4.35±0.02	4.0±0.09	0.92±0.02	0.87±0.02	0.82±0.01	0.94±0.01	4.23±0.01	3.96±0.03	0.94±0.01
High Corruption	0.85±0.01	0.74±0.01	0.87±0.01	4.35±0.02	3.79±0.09	0.87±0.02	0.87±0.02	0.76±0.02	0.87±0.03	4.23±0.01	3.87±0.03	0.91±0.01

	Modality shift (OoD: X-Ray)						Question Type Shift (OoD: Size)					
	Closed Ended			Open Ended			Closed Ended			Open Ended		
	i.i.d.	**OoD**	**RR**	**i.i.d.**	**OoD**	**RR**	**i.i.d.**	**OoD**	**RR**	**i.i.d.**	**OoD**	**RR**
Realistic Shift	0.85±0.01	0.64±0.07	0.75±0.08	4.35±0.02	3.4±0.15	0.78±0.04	0.87±0.02	0.53±0.06	0.61±0.06	4.23±0.01	4.21±0.27	0.99±0.07

E MULTIMODAL SHIFTS

We conducted an ablation study on the OVQA dataset to evaluate the impact of a multimodal shift compared to the previously introduced unimodal shifts. This multimodal shift combines the Manifestation (Body Part) and Question Type Shifts reported in our experiments. Specifically, we defined the OoD set as samples featuring body part ”Leg” and question type ”Organ System”, with all other samples classified as i.i.d. As shown in Figure 13, the multimodal shift demonstrates the lowest robustness compared to unimodal shifts, which is expected given that multimodal shifts represent a more extreme divergence than their unimodal components.

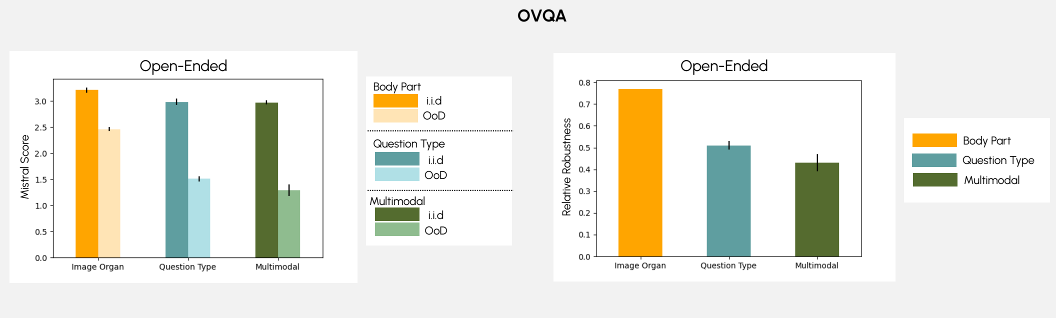


Figure 13: Performance results on OVQA dataset with image organ shift, question type shift and multimodal shift which combines image organ shift and question type shift.