# Exploring the Robustness of Adaptation Methods in VLMs under Realistic Medical Data Shifts

Kim-Celine Kahl*[1,3,7], Selen Erkan*[1,3], Klaus H. Maier-Hein[2,3,4,5,6,7], and Paul F. Jaeger[1,3]

[1] Interactive Machine Learning Group, German Cancer Research Center (DKFZ), Heidelberg, Germany
[2] Division of Medical Image Computing, DKFZ, Heidelberg, Germany
[3] Helmholtz Imaging, DKFZ, Heidelberg, Germany
[4] Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany
[5] National Center for Tumor Diseases (NCT) Heidelberg, Germany
[6] Medical Faculty Heidelberg, University of Heidelberg, Heidelberg, Germany
[7] Faculty of Mathematics and Computer Science, University of Heidelberg, Heidelberg, Germany
{k.kahl, selen.erkan}@dkfz-heidelberg.de

**Abstract.** We evaluate adaptation methods for Vision Language Models under realistic medical data shifts. Our results show that LoRA performs best on the i.i.d set, $(\texttt{IA})^3$ is the most robust, and prompt tuning is insufficient in practice. Further, the question type shift is most severe, and closed-ended questions are less robust than open-ended ones.

**Keywords:** Robustness · VLMs · Adaptation Methods.

*Introduction.* Vision Language Models (VLMs) are increasingly applied in the medical domain, e.g. for Visual Question Answering (VQA). Pretrained medical VLMs, such as [5], require fine-tuning on specific datasets, often using parameter-efficient fine-tuning (PEFT) methods. Further, VLMs must be robust against shifts in the data during application, s.a. unseen diseases, imaging modalities, or question types. Although adaptation methods [8] and their robustness [1] have been benchmarked for general VL problems, this has not been done under realistic and medical shifts. Our study addresses this gap for VQA, and investigates: 1) Which adaptation method yields best performance on i.i.d. data? 2) Which adaption method is most robust? 3) Which shifts most affect model robustness?

*Methods & Experimental Setup.* We finetune the language part of LLaVA-Med [5] using prompt tuning [4], LoRA [2], and $(\texttt{IA})^3$[7]. As VQA datasets, we use SLAKE [6] and OVQA [3], and induce realistic shifts as shown in Table 1, ("OoD:xxx" means unseen image/question type). For evaluation we use Mistral, scoring 0/1 for closed-ended and 1-5 for open-ended questions. The robustness

*These authors contributed equally to this work

**Table 1.** Robustness results. All PEFT method experiments ran for three seeds.

| SLAKE | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 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 FT** | 0.5 | 0.25 | 0.5 | 2.54 | 2.66 | 1.05 | 0.43 | 0.82 | 1.91 | 2.65 | 1.74 | 0.66 |
| **Prompt** | 0.69±0.05 | 0.54±0.04 | **0.78±0.02** | 3.25±0.54 | 2.82±0.2 | **0.88±0.13** | 0.67±0.05 | 0.35±0.06 | 0.52±0.05 | 3.37±0.47 | 3.1±1.12 | 0.9±0.25 |
| **LoRA** | **0.84±0.01** | 0.6±0.09 | 0.71±0.11 | **4.28±0.01** | **3.56±0.01** | 0.83±0.0 | **0.85±0.0** | 0.43±0.09 | 0.51±0.11 | **4.2±0.05** | 3.93±0.1 | 0.94±0.03 |
| **(IA)$^3$** | 0.82±0.0 | **0.63±0.11** | 0.77±0.13 | 4.21±0.02 | 3.35±0.04 | 0.8±0.01 | 0.84±0.01 | **0.49±0.07** | **0.58±0.08** | 4.19±0.04 | **4.02±0.26** | **0.96±0.06** |
| **Most Freq.** | 0.69 | - | - | 3.22 | - | - | 0.696 | - | - | 3.05 | - | - |

| OVQA | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Organ (Location) 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 FT** | 0.42 | 0.47 | 1.12 | 2.34 | 2.43 | 1.04 | 0.43 | 0.57 | 1.34 | 2.34 | 1.89 | 0.81 |
| **Prompt** | 0.64±0.06 | 0.53±0.06 | 0.82±0.03 | 2.62±0.18 | 2.3±0.11 | **0.88±0.08** | 0.56±0.12 | **0.63±0.15** | **1.17±0.41** | 2.46±0.21 | **1.72±0.18** | **0.7±0.1** |
| **LoRA** | **0.84±0.02** | **0.75±0.01** | **0.89±0.01** | **3.18±0.04** | **2.41±0.02** | 0.76±0.01 | **0.81±0.02** | 0.12±0.05 | 0.15±0.06 | **3.01±0.04** | 1.35±0.08 | 0.45±0.03 |
| **(IA)$^3$** | 0.79±0.02 | 0.7±0.01 | **0.89±0.01** | 2.94±0.11 | 2.41±0.05 | 0.82±0.04 | 0.73±0.03 | 0.25±0.12 | 0.34±0.18 | 2.72±0.02 | 1.39±0.03 | 0.51±0.01 |
| **Most Freq.** | 0.75 | - | - | 2.57 | - | - | 0.73 | - | - | 2.23 | - | - |

is measured by the relative robustness as $RR = 1 - ((P_I - P_O)/P_I)$ [1], where $P_I$ is the i.i.d and $P_O$ the OoD performance. As a baseline for the i.i.d set, we selected the most frequent answer to each question that was in the training set.

*Results & Conclusion.* The results in Table 1 show that finetuned models always drop in performance from i.i.d. to OoD, unlike non-finetuned models, showing that the shifts are not inherently more difficult. LoRA, which is the best performing PEFT method on the i.i.d. set, and $(\texttt{IA})^3$ always outperform the baseline, while prompt tuning doesn't always, making it infeasible in practice. Although the robustness trends are more consistent within dataset shifts than within the models, $(\texttt{IA})^3$ often shows slightly higher robustness than LoRA (marked blue). For the severity of the shifts, the closed-ended questions yield a lower RR than the open-ended questions, and the question type shifts seem more severe than the others. The fact that the question type shift has the most severity and the "most frequent" baseline outperforms random performance suggest two research directions: finetuning the vision encoder alongside the language model is likely crucial in the medical domain, and medical VQA datasets often lack question diversity, which can bias models.

## References

1. Chen, S., Gu, J., Han, Z., Ma, Y., Torr, P., Tresp, V.: Benchmarking robustness of adaptation methods on pre-trained vision-language models. In: Advances in Neural Information Processing Systems. vol. 36, pp. 51758–51777 (2023)
2. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models (2021)
3. Huang, Y., Wang, X., Liu, F., Huang, G.: OVQA: A Clinically Generated Visual Question Answering Dataset. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2924–2938. SIGIR '22 (2022)
4. Lester, B., Al-Rfou, R., Constant, N.: The Power of Scale for Parameter-Efficient Prompt Tuning (2021)

5. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In: Advances in Neural Information Processing Systems. vol. 36, pp. 28541–28564 (2023)
6. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: 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 (2021)
7. Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., Raffel, C.A.: Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In: Advances in Neural Information Processing Systems. vol. 35, pp. 1950–1965 (2022)
8. Sung, Y.L., Cho, J., Bansal, M.: 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)