Querying GI Endoscopy Images: A VQA Approach

Notebook for ImageCLEFmedical at CLEF 2025

Gaurav Parajuli[1,*,†]

[1]Johannes Kepler Universität Linz, Austria

Abstract

VQA (Visual Question Answering) combines Natural Language Processing (NLP) with image understanding to answer questions about a given image. It has enormous potential for the development of medical diagnostic AI systems. Such a system can help clinicians diagnose gastro-intestinal (GI) diseases accurately and efficiently. Although many of the multimodal LLMs available today have excellent VQA capabilities in the general domain, they perform very poorly for VQA tasks in specialized domains such as medical imaging. This study is a submission for ImageCLEFmed-MEDVQA-GI 2025 subtask 1 that explores the adaptation of the Florence2 model to answer medical visual questions on GI endoscopy images. We also evaluate the model performance using standard metrics like ROUGE, BLEU and METEOR. The code used in the experiments is publicly available at: github.com/gauravparajuli/ImageCLEFmed-MEDVQA-GI-2025-Task1.

Keywords

Medical VQA, Florence-2, LoRA, ImageCLEFmed 2025, Multimodal AI, Supervised Fine-tuning, Clinical Question Answering, Gastrointestinal Imaging, Kvasir-VQA

1. Introduction

The gastrointestinal (GI) tract consists of all major organs in the digestive system, including the esophagus, stomach, and intestines. The GI tract is vulnerable to a wide range of abnormal mucosal conditions, ranging from minor irritations to highly lethal diseases [1, 2]. Globally, 4.8 million new cases are reported annually, resulting in approximately 3.4 million deaths per year [3], highlighting the high mortality rate associated with GI diseases. Endoscopy is a medical procedure in which a thin device called an endoscope is inserted directly into the body to view organs and other structures. This allows the doctor to diagnose and sometimes treat the condition without surgery. Although endoscopy is a gold-standard procedure, there is about a 20% miss rate for polyp identification in the colon due to operator error [4, 5].

This highlights the need for AI system intervention in medical diagnostics [6]. Medical VQA (MED-VQA) combines natural language processing with image understanding to answer diagnostic questions related to medical images. Such an AI-enabled support system offers promise in helping healthcare professionals provide high-quality care on a large scale accurately and efficiently.

While many available multimodal LLMs perform decently out of the box in the general domain, these same LLMs struggle greatly with VQA tasks in specialized domains such as medical imaging. This raises the need for supervised fine-tuning on a specialized medical dataset. This study is a submission for ImageCLEFmed MEDVQA-GI 2025 subtask 1. It explores how a general-purpose model can be adapted for better performance in specialized domains like medical imaging via supervised fine-tuning.

2. Related Work

Numerous efforts have been made to advance the field of MEDVQA. Since 2018, the annual ImageCLEF MEDVQA benchmark has played a critical role in pushing the field forward. Previously, MEDVQA

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

parajuligaurav007@gmail.com (G. Parajuli)

https://gauravparajuli.github.io/ (G. Parajuli)

0009-0005-5379-3686 (G. Parajuli)

methods combined CNNs and BERT with attention-based fusion for radiology images [7, 8, 9]. In addition, the datasets lacked focus on GI endoscopy. However, with the introduction of the KVASIR [10] dataset and the recently introduced KVASIR-VQA [11], this issue has been mitigated. Current datasets provide question–answer pairs on GI tract images [12]. With the emergence of transformer-based models that have demonstrated strong performance in various deep learning tasks, this work focuses on fine-tuning Florence2 using the KVASIR-VQA dataset to enable effective VQA on GI endoscopic images.

3. Task Overview and Dataset

We participated in subtask 1 (VQA) of the ImageCLEFmed 2025 [13] challenge. The objective of this challenge was to develop a model capable of accurately interpreting and answering all the questions related to GI images.

For this task, we used the KVASIR-VQA dataset. The Kvasir-VQA dataset is a combination of HyperKvasir [14] and KvasirInstrument [15]. It consists of 6500 images across five different categories:

Table 1
Image Category in Kvasir-VQA

Category	Count
Normal	2500
Polyps	1000
Esophagitis	1000
Ulcerative Colitis	1000
Instrument	1000
Total	6500

Each of these 6500 images in KVASIR-VQA is accompanied by various question–answer pairs. There are a total of six question types in the dataset.

1. Yes/No *Example: Does this image contain any finding?*
2. Single choice *Example: What type of polyp is taken?*
3. Multiple choice *Example: Are there any anatomical landmarks in the images?*
4. Choice(Color) *Example: What color is the abnormality?*
5. Location *Example: Where in the image is the abnormality?*
6. Numerical counting *Example: How many polyps are there in the image?*

Since about 58,800 question–answer pairs were available for the 6500 images, no attempt was made to augment the dataset by paraphrasing question–answer pairs.

4. Methodology

4.1. Architecture Overview

In this study, we fine-tuned the Florence2 model from Microsoft on KVASIR-VQA for medical VQA. Florence2 uses prompt-based multitask learning, which enables it to perform a diverse set of tasks such as object detection, segmentation and image captioning without the need for task-specific heads [16]. The Florence2 architecture consists of (a) a vision encoder (DaViT) that converts images into visual token embeddings, (b) a text encoder that processes prompt-style questions, (c) a multimodal transformer encoder–decoder that fuses image and text tokens and (d) a generative output that follows autoregressive decoding schemes like other LLMs.

4.2. Training Details

No augmentations were made to the training dataset provided by the organizer. The dataset was divided (using a seed value) into training and validation sets in a 9:1 ratio. We used LoRA (Low-Rank Adaptation) adapters for fine-tuning. Training was carried out for 5 epochs (early stopping patience set to 2 epochs) with evaluation at the end of each epoch on an NVIDIA RTX A4000 16GB GPU. We used Weights and Biases (wandb) to track training progress.

4.2.1. Hyperparameters Search

In order to determine the optimal hyperparameters, we used Optuna with Bayesian optimization to perform a hyperparameter search. We performed 100 search trials (one epoch each) using only 2.5% of the actual dataset. Across all trials, the seed was fixed for reproducibility. Hyperparameter search ranges for the trials were: (1) learning rate: [1e-6, 1e-4], (2) batch size per device: [2, 4], (3) gradient accumulation steps: [1, 2, 4], (4) weight decay: [0.0, 0.1], (5) LoRA rank: [4, 8, 16], (6) LoRA alpha: [8, 16, 32] and (7) LoRA dropout: [0.0, 0.3]. After 100 search trials, the best configuration found was: (a) learning rate: 9.59e-5, (b) batch size per device: 2, (c) gradient accumulation steps: 2, (d) weight decay: 0.071, (e) LoRA rank: 16, (f) LoRA alpha: 32, (g) LoRA dropout: 0.05478.

However, the initial training with the above hyperparameters on the entire training set led to gradient explosion. Therefore, the learning rate obtained from the search was scaled down by a factor of four (referred to as the base learning rate from now on). Finally, to reduce the noise in the training loss curve, the effective batch size was increased to 64 via gradient accumulation and the base learning rate was scaled as:

$$\sqrt{\frac{\text{new\_batch\_size}}{\text{old\_batch\_size}}}$$

Rest of the hyperparameters were kept as per the hyperparameters search result.

5. Results and Evaluation

The following table outlines the performance of our best model on both the public and private test sets of the organizing team.

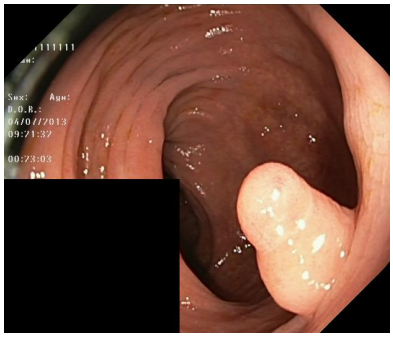
Table 2
Evaluation Metrics for Test Set

Test Set	BLEU	ROUGE1	ROUGE2	ROUGEL	METEOR
Public	0.21	0.87	0.12	0.86	0.48
Private	0.18	0.91	0.11	0.9	0.5

The following table summarizes the performance of our best model across different question types.

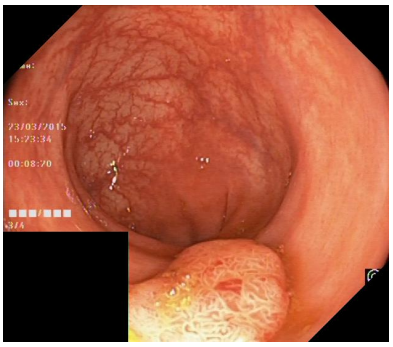
Table 3
Evaluation Metrics for Test Set(across different question types)

Question Type	BLEU	ROUGE1	ROUGE2	ROUGEL	METEOR
YesNo	0.0	0.95	0.14	0.95	0.54
Choice (Singular)	0.0	0.8	0.08	0.8	0.42
Choice (Multiple)	0.0	0.84	0.01	0.84	0.42
Choice (Color)	0.47	0.78	0.32	0.73	0.51
Location	0.0	0.84	0.21	0.81	0.47
Numerical Count	0.0	0.9	0.0	0.9	0.45
Overall	0.21	0.87	0.12	0.86	0.48



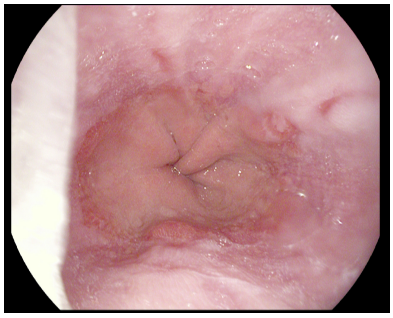
Question: Where in the image is the abnormality?
Prediction: center; center-right; lower-right
Ground Truth: center; center-right; lower-center; lower-right

(a) VQA on GI Endoscopy Image (Location Question)



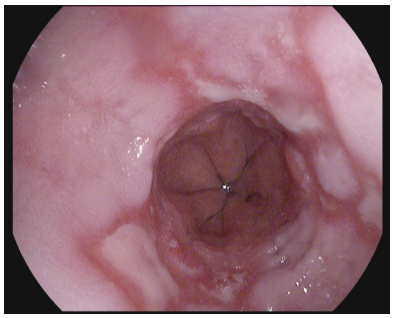
Question: Are there any abnormalities in the image? Check all that are present.
Prediction: polyp
Ground Truth: polyp

(b) VQA on GI Endoscopy Image (Multiple Choice Question)



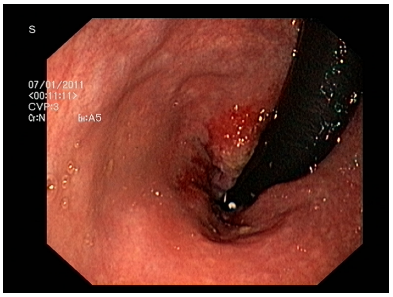
Question: How many polyps are in the image?
Prediction: 0
Ground Truth: 0

(c) VQA on GI Endoscopy Image (Numerical Counting Question)



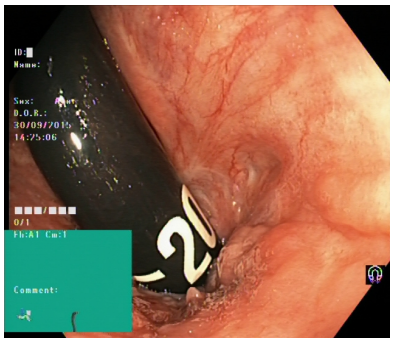
Question: What color is the abnormality? If more than one separate with ;
Prediction: pink; red; white
Ground Truth: pink; red; white

(d) VQA on GI Endoscopy Image (Choice Color Question)



Question: Is there a green/black box artefact?
Prediction: no
Ground Truth: no

(e) VQA on GI Endoscopy Image (Yes/No Question)



Question: What type of procedure is the image taken from?
Prediction: colonoscopy
Ground Truth: colonoscopy

(f) VQA on GI Endoscopy Image (Single Choice Question)

Figure 1: VQA on GI Endoscopy Images for different question types

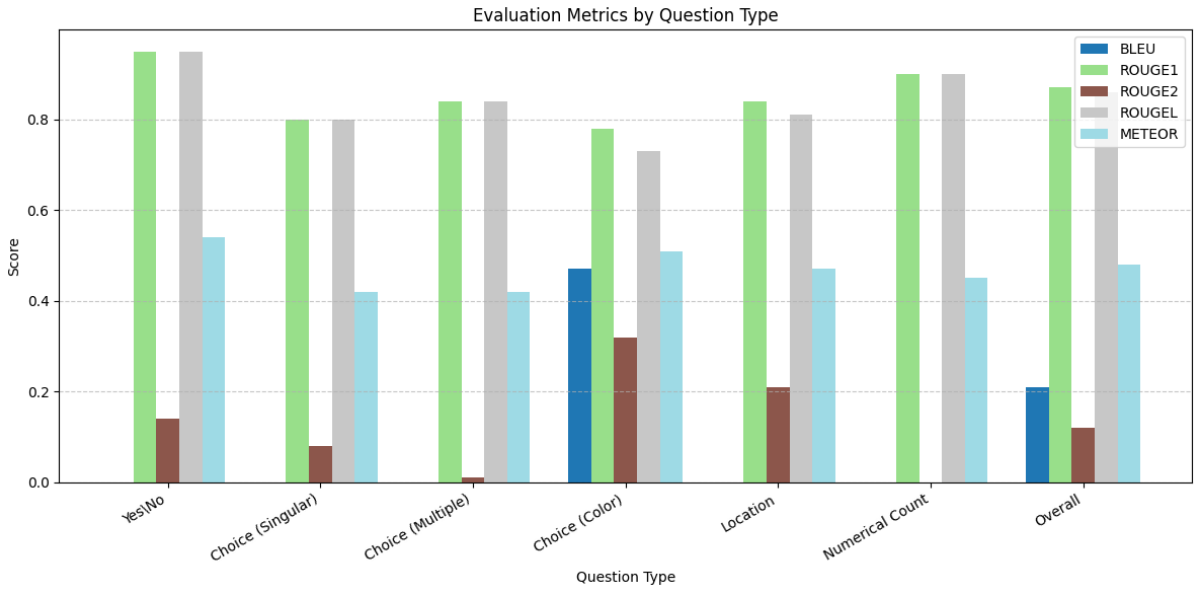


Figure 2: Evaluation Metrics across different question types

6. Ablation Studies

To understand the impact of different fine-tuning strategies and the role of different components, we performed the following ablation studies using four variants of the Florence2 model.

Table 4
Ablation studies results for GI MEDVQA

Model Variant	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
florence2_64_FT	0.2	0.85	0.11	0.85	0.47
florence2_64_FT_EF	0.17	0.84	0.1	0.84	0.46
Florence2_64_r8	0.23	0.85	0.11	0.84	0.46
Florence2_64_r16	0.21	0.87	0.12	0.86	0.48

Please note that the model variant is in the format `baseModelName_batchSize_trainingStrategy`.
Model weights for the above variants are available at: *https://www.hf.co/gauravparajuli/model_variant_name*

For the first variant, we froze the vision tower in the Florence2 architecture and proceeded with training. For the second variant, we froze the encoder portion of the Florence2 language model in addition to the vision tower. For the third and fourth variants, we used LoRA adapters with rank=8, alpha=16 and rank=16, alpha=32 respectively.

Here we can clearly see that the LoRA-based variant with rank=16 outperforms all variants. The LoRA-based variant with rank=8 achieves the highest BLEU score, possibly due to a regularization effect from the reduced parameter count, but it slightly underperforms on all other metrics.

The second variant, in which both the encoder and the vision tower were frozen, performs slightly under the first variant, in which only the vision tower was frozen. This suggests that while fine-tuning the decoder alone can capture some useful adaptation, the encoder's contribution is crucial for optimal performance in the multimodal task.

In general, these results validate the effectiveness of LoRA-based fine-tuning on downstream tasks.

7. Discussion and Comparison with Literature

Our approach outshines the performance of the previous year's baseline, which had a ROUGE1 score of 0.6955. However, despite achieving a strong ROUGE score, our model performed worse than the previous year's baseline in terms of the BLEU score (0.21 vs. 0.3757). The previous baseline was trained on only 2000 images (a small subset). As the BLEU metric penalizes short candidates, it is plausible that our model, which possibly generated more diverse and concise answers due to greater data exposure, was disproportionately penalized.

Also, Kvasir-VQA contains several question types. For "Yes/no" question and single word answer question, higher ROUGE score is easier to achieve as the vocabulary is limited. However, the BLEU score in this case will be very low or zero if the prediction does not exactly match the single word ground truth. This is the reason why the BLEU score was zero for the majority of question types in Table 3.

8. Conclusion and Future Work

It is evident from this study that it is possible to train a robust MEDVQA system based on Kvasir-VQA. Notable future directions for this study include:
1. Expanding the current dataset by introducing more diverse images and question answer pairs. This might help in developing more robust and generalizable models.
2. Increase the performance of the model on the BLEU metric without sacrificing the performance on the ROUGE metric. This could involve exploring different decoding strategies and loss functions that encourage more grammatically correct sentences.

3. Extensive evaluation of the model in real world clinical settings to gauge potential for real world applications.

Acknowledgments

We would like to thank the organizers from the SimulaMet Department of Holistic Learning who made this event possible. This work heavily relied on the KVASIR-VQA dataset, which was also compiled by SimulaMet, and we deeply appreciate their contribution. Additionally, we would like to thank the researchers from Microsoft for their Florence-2 model. Lastly, this work would not have been possible without Hugging Face. We heartily thank everyone who has contributed to the Transformers library and the PEFT library within the Hugging Face ecosystem.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT for grammar and spelling checks. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] H. Gelberg, Pathophysiological mechanisms of gastrointestinal toxicity, Comprehensive Toxicology (2017) 139.

[2] C. B. Navarre, D. Pugh, Diseases of the gastrointestinal system, Sheep & Goat Medicine (2009) 69.

[3] M. Arnold, C. C. Abnet, R. E. Neale, J. Vignat, E. L. Giovannucci, K. A. McGlynn, F. Bray, Global burden of 5 major types of gastrointestinal cancer, Gastroenterology 159 (2020) 335–349.

[4] O. F. Ahmad, A. S. Soares, E. Mazomenos, P. Brandao, R. Vega, E. Seward, D. Stoyanov, M. Chand, L. B. Lovat, Artificial intelligence and computer-aided diagnosis in colonoscopy: current evidence and future directions, The lancet Gastroenterology & hepatology 4 (2019) 71–80.

[5] P. Wang, T. M. Berzin, J. R. G. Brown, S. Bharadwaj, A. Becq, X. Xiao, P. Liu, L. Li, Y. Song, D. Zhang, et al., Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study, Gut 68 (2019) 1813–1819.

[6] M. Chaichuk, S. Gautam, S. Hicks, E. Tutubalina, Prompt to Polyp: Medical Text-Conditioned Image Synthesis with Diffusion Models, arXiv (2025). doi:10.48550/arXiv.2505.05573. arXiv:2505.05573.

[7] X. Yan, L. Li, C. Xie, J. Xiao, L. Gu, Zhejiang university at imageclef 2019 visual question answering in the medical domain, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, 2019. URL: https://www.imageclef.org/2019/medical/vqa, team Hanlin.

[8] A. Ben Abacha, M. Sarrouti, D. Demner-Fushman, S. A. Hasan, H. Müller, Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain, in: CEUR Workshop Proceedings, volume 2936, 2021. URL: http://ceur-ws.org/Vol-2936.

[9] A. Ben Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, H. Müller, Vqa-med: Overview of the medical visual question answering task at imageclef 2019, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, 2019. URL: https://github.com/abachaa/VQA-Med-2019.

[10] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, et al., Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection, in: Proceedings of the 8th ACM on Multimedia Systems Conference, 2017, pp. 164–169.

[11] S. Gautam, A. M. Storås, C. Midoglu, S. A. Hicks, V. Thambawita, P. Halvorsen, M. A. Riegler, Kvasir-VQA: A Text-Image Pair GI Tract Dataset, in: ACM Conferences, Association for Computing Machinery, New York, NY, USA, 2024, pp. 3–12. doi:10.1145/3689096.3689458.

[12] S. Gautam, M. A. Riegler, P. Halvorsen, Kvasir-VQA-x1: A Multimodal Dataset for Medical Reasoning and Robust MedVQA in Gastrointestinal Endoscopy, arXiv (2025). doi:10.48550/arXiv.2506.09958. arXiv:2506.09958.

[13] B. Ionescu, H. Müller, D.-C. Stanciu, A. Idrissi-Yaghir, A. Radzhabov, A. G. S. de Herrera, A. Andrei, A. Storås, A. B. Abacha, B. Bracke, B. Lecouteux, B. Stein, C. Macaire, C. M. Friedrich, C. S. Schmidt, D. Fabre, D. Schwab, D. Dimitrov, E. Esperança-Rodier, G. Constantin, H. Becker, H. Damm, H. Schäfer, I. Rodkin, I. Koychev, J. Kiesel, J. Rückert, J. Malvehy, L.-D. Ștefan, L. Bloch, M. Potthast, M. Heinrich, M. A. Riegler, M. Dogariu, N. Codella, P. H. P. Nakov, R. Brüngel, R. A. Novoa, R. J. Das, S. A. Hicks, S. Gautam, T. M. G. Pakull, V. Thambawita, V. Kovalev, W.-W. Yim, Z. Xie, Overview of imageclef 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.

[14] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, et al., HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy, Sci. Data 7 (2020) 1–14. doi:10.1038/s41597-020-00622-y.

[15] D. Jha, S. Ali, K. Emanuelsen, S. A. Hicks, V. Thambawita, E. Garcia-Ceja, M. A. Riegler, T. De Lange, P. T. Schmidt, H. D. Johansen, et al., Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy, in: MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27, Springer, 2021, pp. 218–229.

[16] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, L. Yuan, Florence-2: Advancing a unified representation for a variety of vision tasks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 4818–4829.