Generative AI Based on Medical Visual Question Answering (VQA) Techniques

by
Sarthak Kaushik

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science in Computer Science (MSc)

in the
Department of Computer Science
Faculty of Computer Science

LAKEHEAD UNIVERSITY
April 2026

Declaration of Committee

Name:	**Sarthak Kaushik**
Degree:	**Master of Science in Computer Science (MSc)**
Title:	**Generative AI Based on Medical Visual Question Answering (VQA) Techniques**
Committee:	**Chair: Dr. Sabah Mohammed** Supervisor Chair, Department of Computer Science
	Dr. Jinan Fiaidhi Supervisor Full Professor, Department of Computer Science
External Evaluator:	**Dr. Carlos Zerpa** Associate Professor, Department of Kinesiology Lakehead University

Ethics Statement

The current dissertation is founded on the secondary analysis of the available datasets concerning gastrointestinal endoscopy and ulcerative colitis severity determination. Nobody was recruited in this study and no new clinical data were obtained. This statement will be reviewed and changed accordingly (where necessary) after the institution has made a decision on the Research Ethics Board approval or exemption.

Abstract

Medical visual question answering (MedVQA) enables clinicians to pose direct medical image questions rather than just using single-label image classification. It comes in handy in gastrointestinal (GI) endoscopy and, more specifically, in colonoscopy where clinical inquiries tend to be about whether findings are present, where they are, how many they are, or how much disease they represent. Nevertheless, the existing GI MedVQA systems are still plagued with critical issues, such as unbalanced dataset structure, imbalanced classes, mismatch of the answer format, and poor visual grounding in case of free-text generation.

These problems are examined in this dissertation based on five GI datasets: HyperKvasir, Kvasir-VQA, Kvasir-VQA-x1, ImageCLEF MEDVQA-GI and LIMUC. The comparison of benchmarking results reveals a similar trend in these datasets: in case of current GI tasks, pipelines with supervision or other restrictions are more trustworthy than unsupervised and raw zero-shot vision-language generation. This is particularly relevant in the context where the answer space is fixed or where the clinically important classes are uncommon.

The dissertation develops a generative approach to ulcerative colitis severity scoring on LIMUC based on this finding in which the Mayo endoscopic subscore is the target task. The model is tested to explicit output limitations, and analysis with clinically relevant analysis. The work then maps this model-level contribution to a physician-support environment by a modular wrapper which interprets queries, retrieves supporting evidence as necessary, generates citation-linked answers, rejects unsupported queries, and maintains traceable system logs.

The general argument of the dissertation is that GI MedVQA can only proceed toward clinical decision support with systems that are still visually based, whose limits can be traced, and whose auditing can be performed by the supervision of physicians. The work does not purport independent clinical application. Rather, it provides a gradual route to benchmark assessment to safer clinician-facing colonoscopy aid.

Keywords medical visual question answering, gastrointestinal endoscopy, colonoscopy, ulcerative colitis, Mayo endoscopic subscore, vision-language models, retrieval-augmented generation, clinical decision support.

Dedication

To my family, professors, mentors, friends, and all those who supported me throughout this journey. This work is dedicated to all of you.

Acknowledgements

This dissertation would not have been possible without the support, guidance, and encouragement of many people, and I am deeply grateful to all of them.

Above all, I thank my family for their constant love, patience, sacrifice, and belief in me. Their encouragement sustained me throughout this journey and provided the foundation on which this achievement rests.

I am sincerely grateful to Dr. Sabah Mohammed and Dr. Jinan Fiaidhi for their guidance, academic support, and valuable insight throughout this research. Their encouragement and scholarly direction shaped this dissertation and contributed greatly to my growth as a researcher.

I am also deeply grateful to Dr. Sabah Mohammed for supporting my involvement in the MITACS Accelerate initiative. Through this opportunity, I was able to collaborate with Aurora Constellations under the guidance of Dr. Arnold Kim, which gave me valuable exposure to the industrial application of related research.

I am also thankful to my professors and mentors, whose teaching, advice, and example have influenced my academic journey in lasting and meaningful ways.

I would like to offer special thanks to Dr. Arnold Kim for his generosity, guidance, and encouragement. His support and perspective have meant a great deal to me.

My thanks also go to my friends for their understanding, companionship, and encouragement throughout this demanding journey. Their support made difficult moments easier and important milestones more meaningful.

Finally, I would like to acknowledge everyone who contributed to this work, directly or indirectly. This dissertation reflects not only my own effort, but also the kindness, confidence, and support of the many people who stood by me along the way.

Table of Contents

List of Acronyms and Abbreviations

Acronym	Full Form
AI	Artificial Intelligence
ANLS	Average Normalized Levenshtein Similarity
BBPS	Boston Bowel Preparation Scale
BioBERT	Biomedical Bidirectional Encoder Representations from Transformers
BioGPT	Biomedical Generative Pre-trained Transformer
BLEU	Bilingual Evaluation Understudy
BLIP	Bootstrapping Language-Image Pre-training
CAD	Computer-Aided Diagnosis
CDS	Clinical Decision Support
CI	Confidence Interval
CIDEr	Consensus-based Image Description Evaluation
CLIP	Contrastive Language-Image Pre-training
CNN	Convolutional Neural Network
CV	Computer Vision
ECE	Expected Calibration Error
EM	Exact Match
F1	F1-score
GenAI	Generative Artificial Intelligence
GI	Gastrointestinal
GRU	Gated Recurrent Unit
KB	Knowledge Base
LIMUC	Labeled Images for Ulcerative Colitis
LLaVA	Large Language and Vision Assistant
LLM	Large Language Model
LoRA	Low-Rank Adaptation
LVLM	Large Vision-Language Model
LXMERT	Learning Cross-Modality Encoder Representations from Transformers
MAE	Mean Absolute Error
MCC	Matthews Correlation Coefficient
MedVQA	Medical Visual Question Answering
MedVQA-GI	Medical Visual Question Answering for Gastrointestinal Tract
MES	Mayo Endoscopic Subscore
METEOR	Metric for Evaluation of Translation with Explicit Ordering
MLLM	Multimodal Large Language Model
NLP	Natural Language Processing
NLG	Natural Language Generation
OOV	Out-of-Vocabulary
PICO	Patient/Population, Intervention, Comparator/Comparison, Outcome(s)

QA	Question Answering
QWK	Quadratic Weighted Kappa
RAG	Retrieval-Augmented Generation
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
ROUGE-L	Recall-Oriented Understudy for Gisting Evaluation – Longest Common Subsequence
RQ	Research Question
TF-IDF	Term Frequency–Inverse Document Frequency
UC	Ulcerative Colitis
UCEIS	Ulcerative Colitis Endoscopic Index of Severity
ViLBERT	Vision-and-Language BERT
ViLT	Vision-and-Language Transformer
ViT	Vision Transformer
VLQA	Visual Location Question Answering
VLM	Vision-Language Model
VQA	Visual Question Answering
VQG	Visual Question Generation

Chapter 1.

Introduction

1.0 Chapter 1 Overview

Chapter 1 describes the entire research background of this dissertation and explains why GI-endoscopy MedVQA should not be assessed using the aggregate benchmark scores. The chapter initially defines the clinical rationale behind the use of question-driven AI assistance in the context of colonoscopy, followed by formalizing the key methodological gap: existing systems can generate high average-performing results and still fail on a clinically high-risk, imbalanced, or format-sensitive case. It then achieves this gap into a research program through the statement of the thesis aim, objectives, research questions and working hypotheses. Moreover, Chapter 1 establishes scope of operation of the work, what is explicitly out of scope, and how PICO-oriented evidence mapping can be a controlled extension pathway and not a substitution of visual reliability. Lastly, the chapter presents the framing scenario-driven, conceptual pipeline, and risk-control logic that would inform all the experiments and interpretations to come.

Collectively, Chapter 1 gives context and rationale behind MedVQA in GI endoscopy, the reliability and translational gap that this dissertation will address, the study aim and research questions as per the evaluation method, the scope and assumptions under which the claims being made are limited, and the roadmap of Chapters 2 and 3. Chapter 2 expands on this basis with a systematic scoping review which determines the most applicable model families, benchmark practices, and open gaps to GI-oriented evaluation.

1.1 Background & Motivation

Medical Visual Question Answering (MedVQA) is a field that is at the crossroads of computer vision (CV), natural language processing (NLP), and clinical medicine. It aims to create artificial intelligence that will be able to read medical images and respond to clinically significant questions in natural language. Contrary to traditional image classification systems, which tend to provide only categorical responses, MedVQA is based on question interaction. This form of interaction is more aligned with clinician thinking in practice where interpretation is often informed by targeted questions as opposed to discrete labels.

MedVQA has a great potential in clinical decision support. It could enhance consistency in diagnosis, lessen physician workloads, and enhance access to specialist-level interpretation during telemedicine and resource-constrained environments. Its long-term aim is not to substitute clinical judgment, but to act as a smart and open assistant, which assists clinical reasoning in the current workflows.

MedVQA has domain constraints as compared to general-domain VQA. Medical images usually have subtle, overlapping, or low-contrast findings which necessitate specialized interpretation. There can also be clinical questions with multi-step reasoning, and visual evidence should be taken into account along with bio-medical context and intent to do something. These characteristics render MedVQA a multimodal task that is technically challenging and a clinically sensitive reliability problem.

MedVQA is based on the technical background of a combination of CV-based image representation and NLP-based question understanding and answer generation. The early systems were mainly discriminative where they were based on an image encoder, a question encoder, fusion module and hard set of answers. Whereas these systems worked reasonably well on more limited tasks, they demonstrated less adaptability to more clinical queries that were more subtle and open ended (Ben Abacha et al. 2019; Borgli et al. 2020; Liu et al. 2021). More recent work has focused on multimodal architectures and generative response models based on transformers, facilitated by large-scale visual instruction tuning and medical vision-language modeling (Zhang et al. 2023; C. Li et al. 2023; Dong et al. 2025) and by modern surveys and benchmarks of multimodal model behavior in MedVQA (Lin et al. 2023; Hu et al. 2024). Meanwhile, biomedical language models, such as BioBERT and BioGPT, have enhanced domain-specific language comprehension and generation (Lee et al. 2019; Luo et al. 2022).

Despite this advancement, a significant disparity between benchmark performance and solid clinical utility still exists. The modern scene is marked by three ongoing tensions. First, generative models are more expressive, but can exhibit lexical drift, ungroundedness, or hallucinated text. Second, minorities may have clinically significant failures that are concealed by high aggregate accuracy. Third, general vision-language models often perform less well in specialized clinical settings in zero-shot transfer. These tensions are the impetus behind this dissertation and they are explored within the context of gastrointestinal endoscopy, and specifically colonoscopy-oriented MedVQA.

1.1.1 From Fixed Answer Systems to Generative Multimodal Models

The MedVQA historical development offers valuable background to this thesis. In previous systems, generation of answers was in effect dealt as label choice. This formu-

lation proved to be fairly easy to optimize and evaluate, but restricted clinically meaningful expression. Consequently, intricate queries were frequently condensed into rough categorical results, diminishing interpretability and restricting usefulness to clinicians.

The transition to transformer-based multimodal models enhanced cross-modal alignment and generated free-text responses. This shift dealt with some significant shortcomings of the previous systems, including the insufficient interaction of questions and images, and the limited scope of predetermined output vocabularies. Simultaneously, it brought with it novel threats. In a clinical situation, a fluency response with a weak grounding in image evidence can be worse than a confined answer the grounding of which can be proved. That is why, model expressiveness should be weighted against grounding, calibration and error control.

Another change has occurred at the evaluation level. Instead of reporting performance on a single benchmark, recent research is looking more at cross-dataset behavior and resilience to distributional change. This is particularly true in GI endoscopy, where there may be a large degree of heterogeneity in questions templates, answer conventions, and visual attributes.

1.1.2 Limitations of Aggregate Benchmark Scores in Clinical AI

Benchmark metrics are still needed but not sufficient in their own right to be clinically validated. One model could have a high total accuracy and still not work on infrequent categories with clinical implications. Underrepresented classes may be masked by the use of weighted scores and single summary measures may not be relevant to the clinical importance of specific classes of errors.

In tasks that are severity-based, the ordinal structure does count as well. Mixing up neighboring grades is not synonymous with the absence of severe disease. In clinical MedVQA evaluation, class-wise measures, measures that take into account class imbalance, ordinally sensitive measures, and scenario-level measures of acceptability should thus be evaluated.

This wider perspective of evaluation is taken in this dissertation starting with Chapter 1. It also considers answer format as a design option which too must be appraised. A label-only output can be adequate in certain environments; in others, clinicians might need a justification, confidence measure, or references to back-up evidence. A good clinical assistant should thus not be evaluated on the basis of predictive accuracy alone but the usefulness of its outputs.

1.2 Why GI Endoscopy and Colonoscopy

Endoscopy of the gastrointestinal (GI) tract is a significant clinical and methodological area of application of MedVQA. Clinically, colonoscopy is important in screening, diagnosing and follow-up in a variety of disease settings, such as colorectal neoplasia and inflammatory bowel disease. The meaning of endoscopic findings is directly related to the escalation of treatment, the planning of surveillance and report on procedures. Focused questions which may be asked by clinicians in routine practice include:

- Does it have an abnormality ?
- What is the number of findings that is visible?
- Where is the lesion located?
- What is the type or morphology?
- How severe are the chances of inflamation?

These types of questions are natural mappings to MedVQA task families, and can be systematically evaluated.

Technically, GI endoscopy is a challenging environment to MedVQA. The quality of the images can be different due to motion blur, specular reflections, debris, variations in bowel preparation and alterations in viewpoint. There is also a tendency to have highly disproportionate question-answer distributions, with the severe classes that have the highest clinical significance often underrepresented. Consequently, colonoscopy-based MedVQA offers a realistic and rigorous environment to assess reliability in domain shift and in class imbalance.

This direction of the thesis is possible due to recent advances in the creation of datasets:

- GI-focused visual question answering at scale (58,849 QA pairs across 6,500 images) was proposed by Kvasir-VQA (Gautam et al. 2024; Simula Datasets n.d.).
- Kvasir-VQA-x1 increased complexity and scale (159,549 QA pairs) and provided transformed/robustness-oriented settings (Gautam et al. 2025a; Simula n.d.).
- ImageCLEF MEDVQA-GI gave shared tracks to GI VQA (Murtaza et al. 2023; ImageCLEF 2023).
- HyperKvasir and LIMUC offer more comprehensive GI visual context and severity-oriented subsets (Borgli et al. 2020; Polat et al. 2022).

The resources can facilitate reproducible analysis, yet they do not eliminate the main task: to develop systems that will be reliable not only to the typical benchmark patterns, but also to clinically high-risk cases.

1.2.1 Colonoscopy Question Families and Their Decision Roles

To maintain methodological clarity, this thesis links each question family to the clinical decision role it is intended to support.

Question family	Example	Dominant challenge	Decision role
Presence	"Is there active bleeding?"	False-negative control	Immediate risk awareness
Count	"How many polyps are visible?"	Counting under occlusion and partial views	Completeness and burden estimation
Location	"Where is the lesion?"	Spatial grounding consistency	Procedural precision and reporting
Type/Attribute	"What polyp morphology is present?"	Fine-grained differentiation	Risk stratification and follow-up planning
Severity	"What is the likely Mayo score?"	Ordinal robustness under imbalance	Treatment planning and monitoring

Table 1.1. Colonoscopy Question Families and Their Decision Roles

1.2.2 Constraints in Real-world Deployment.

Operational constraints must be considered in model design even during pre-deployment research. These are variability in image quality, scarcity of labels in extreme classes, inter-institutional differences in reporting language and the requirement of predictable failure behavior. The difference between a system that is clinically acceptable and one that is never-fail is that the former has failures that are detectable, limited, and remediable.

These considerations justify the use of a conditional architecture. Closed-set pathways can be more suitable where it is necessary to have deterministic control over labels. Explanation-oriented outputs can be generated through generative pathways, although clear protection measures are necessary. The introduction of retrieval-backed reasoning ought to be selective to the questions that cannot be answered by direct visual interpretation.

1.3 Clinical Motivation: Ulcerative Colitis Severity as a Flagship Use Case

High-value use of MedVQA during colonoscopy is ulcerative colitis (UC) severity assessment. Endoscopic severity is a significant factor in the treatment and disease follow-up. Standardized severity scales are usually used to denote severity in clinical practice, like the Mayo Endoscopic Subscore (0-3) and UCEIS (0-8, commonly mapped to

ordinal severity scales). Though these scales are clinically significant, grading may be challenging along the borderline cases or in frames of low image quality.

The idea of automating the severity assessment of UC is not new. Previous deep learning systems have already shown clinically meaningful behavior in endoscopic severity scoring when compared to a specialist in a clinical setting (Lin et al. 2021; Lin et al. 2023; Liu et al. 2021; Liu et al. 2023). Similar studies on the use of GI quality grading with vision-language models also confirm the viability of multimodal assistance in endoscopy procedures (Jiang et al. 2025). The combination of these studies gives solid previous evidence that AI-based severity scoring is indeed a feasible tool in UC, and demonstrates that the next big question is not whether AI can score disease at all, but whether an interactive system can be reliable, well grounded, and clinically useful once it goes beyond fixed-score results (ImageCLEF 2023; ImageCLEF 2025).

In this regard, this dissertation does not purport to make the first AI system in UC severity scoring. Its value-added capability is in (1) its GI-oriented MedVQA interaction, (2) risk-considering reliability controls and (3) management-oriented questions evidence-aware extensions. Regarding a MedVQA point of view, the UC severity is methodologically informative since it is a combination of:

- Fine-grained visual discrimination.
- Ordinal class structure.
- Asymmetric clinical cost of errors.
- Possible need to explain and communicate about uncertainty.

A typical classifier is capable of returning a severity label, but a MedVQA interface may also provide rationale-providing responses and guided follow-up. This thesis takes such a broader view though still possesses a strong focus on visual grounding.

1.3.1 Workflow-Oriented Motivation

In practice, clinical use can be modeled as a sequence: frame(s) of endoscopy, clinician query, answer grounded in visuals, confidence or uncertainty indicator, and, when necessary, evidence-based follow-up. The thesis is designed using this abstraction. MedVQA is considered as a perception task, a language task, and a workflow integration task in this work.

1.3.2 Why UC Severity Is a High-Value Testbed for Chapter 1

Chapter 1 has employed UC severity as a flagship testbed since it combines a number of difficult properties into one clinically relevant problem: subtle visual differences, class imbalance, ordinal grading, and the grave impact of under-calling severe disease. In this context, a system that is not stable is not very suitable in the wider aspect of GI decision support.

Meanwhile, UC severity is experimentally accessible, using existing datasets and preserved repository artifacts. This clinical significance and methodological tractability are why it is a good anchor of this research.

1.4 Problem Statement

The key question that is going to be answered in this dissertation is:

What does a colonoscopy-oriented MedVQA pipeline need to do to go beyond benchmark level answering to evidence-based, reliable and clinically grounded decision support?

This question is also explored in terms of four gaps in operation:

- Data and coverage gap: The current datasets offer a broad range of families of tasks but clinical depth and long-tail bias.
- Model gap: Zero-shot general VLMs often perform poorly compared to constrained or supervised methods in GI tasks.
- Evaluation gap: Aggregate measures may mask patterns of errors that are not acceptable clinically.
- Workflow gap: Numerous systems are currently not linked with image-grounded outputs to evidence-driven clinical reasoning.

1.4.1 Evidence from This Repository

Table 1.2 summarizes key empirical signals from persisted repository artifacts.

References included in the Source column in the below table refers to supplementary analysis artifacts in the thesis repository and are included alongside the dissertation as reproducibility materials.

Dataset / Task	Strongest saved model result	Key failure signal	Source

ImageCLEF MEDVQA-GI 2023 validation	ViLT accuracy 0.9089, macro-F1 0.5823	Qwen2.5-VL zero-shot raw accuracy 0.0007; projected 0.0670	Appendix A, Artifact A2
HyperKvasir 23-class test	ResNet50 supervised accuracy 0.8789, macro-F1 0.5943	Head-tail recall gap remains large; BLIP2 projected accuracy 0.0638	Appendix A, Artifact A1
LIMUC Mayo severity test	Fine-tuned ResNet50 accuracy 0.7539, macro-F1 0.6829, QWK 0.8351	Zero-shot VLM macro-F1 0.1771, balanced accuracy 0.25	Appendix A, Artifact A3
Kvasir-VQA yes/no subset	ResNet+GRU accuracy 0.9865, macro-F1 0.9650	Free-generation run shows unknown-rate collapse in persisted artifact	Appendix A, Artifact A4
Kvasir-VQA-x1 generative track	MedGemma LoRA token-F1 0.5085 (adaptation gain)	Exact-match remains near zero across persisted modern VLM runs	Appendix A, Artifact A5

Table 1.2. Repository Evidence Signals Motivating the Dissertation

In Table 1.2, raw accuracy is the direct label-space scoring with canonical benchmark targets before lexical post-processing. Projected accuracy is a secondary diagnostic where generated free-text answers are deterministically transformed into the known answer space of the individual question and then scored according to the local repository evaluation approach (Appendix A, Artifacts A1–A5).

To reduce the threat of leakage and prevent exaggerated claims, this thesis reports split context and scoring semantics at the dataset level. Official or persisted holdouts are maintained, paired tests are used only on aligned rows where available, and projected scores are interpreted alongside raw exact-match and unknown-output behavior rather than replacing primary metrics (Appendix A, Artifacts A1–A5).

The stage approach is supported by the findings:

- Maintain strong supervised/closed-set visual grounding.
- Add controlled generative ability where it enhances clinical interaction value.
- Introduce retrieval-backed reasoning as a more secure extension within the higher-level questions.

1.4.2 Why This Gap Persists

This disconnect exists since GI MedVQA is not one modeling problem. It involves trusted visual sensory perception in the noisy environment, clinically consistent knowledge of questions, grounded answer production, and safe actions in the environment where the risks are asymmetric. Enhancement of a component does not in all cases lead to end-to-end clinical utility.

A system can be fluent and weakly grounded, is good on common classes and bad in severe cases of minority, or can be stable on fixed prompts but unreliable on linguistic variation. This is the reason why this dissertation considers architecture design, evaluation design and scenario design as interdependent issues and not independent issues.

1.4.3 Consequences of Leaving the Gap Unresolved

Without filling this gap, systems can be perceived to perform well across aggregate benchmarks but still be underperforming in the clinical situations that matter most. This can lead to under-detection of high-risk cases in severity-based settings. In report-based environments, it can result in plausible yet poorly-founded products that contribute to cognitive load. In decision-support applications, it can yield recommendations that are not adequately evidence-based.

These risks defend a conservative methodological stance: reliability, grounding, and traceability must be promoted until output fluency is maximized.

1.5 Aim and Objectives

1.5.1 Aim

This research seeks to design, assess, and report a clinically relevant MedVQA architecture of colonoscopy that integrates sound visual grounding, query-aware generation, and an evidence-aware extension pathway of more risky queries.

1.5.2 Objectives

- Create a reproducible, multi-dataset GI MedVQA benchmark layer with persisted repository artifacts.
- Measure relative reliability of closed-set versus generative methods in GI-specific conditions.
- Assess UC severity-oriented question answering using clinically meaningful measures, such as imbalance-conscious and ordinal-aligned measures.

- Establish scenario motivated use cases that are in line with endoscopist work-flows.

- Definite and experiment with a pathway of retrieval-enhanced, evidence-knowledge answering without interfering with visual-grounded performance.

Objective	Dissertation deliverable
O1	Dataset/task profiling and consolidated benchmark tables
O2	Comparative model analysis with failure-mode interpretation
O3	Severity-focused evaluation section with class-wise and ordinal metrics
O4	Scenario catalogue with acceptance criteria and error taxonomy
O5	Proposed architecture and phased validation plan for evidence-aware extension

Table 1.3. Objective-to-Deliverable Map

1.6 Research Questions

The six research questions guiding this dissertation relate task design, modeling strat-egy, and clinical utility.

RQ1 (Coverage):

What are the currently existing clinically relevant colonoscopy question families and an-swer spaces; what gaps are important?

RQ2 (Comparative reliability):

Are constrained or discriminative pipelines in GI MedVQA tasks more reliable than zero-shot open-ended VLM generation?

RQ3 (Failure modes):

What are the most effectively failure modes in current GI MedVQA systems: class imbal-ance, lexical mismatch, localization ambiguity or domain shift?

RQ4 (Severity robustness):

To what extent can models respond to UC severity-oriented questions, especially un-derrepresented severe classes?

RQ5 (Clinical output format):

What would be the most clinician trusting and usable output format: label only, label with rationale, label with confidence, or answer with retrieved evidence?

RQ6 (Evidence-aware extension):

Is retrieval-augmented reasoning a controllable extension that does not impair the basic visual grounded accuracy?

1.6.1 RQ-to-Evaluation Map

RQ	Primary evidence in this thesis	Candidate metrics
RQ1	Dataset profiling and taxonomy analysis	Coverage %, answer cardinality, imbalance ratio
RQ2	Closed-set vs generative comparisons	Accuracy, macro-F1, MCC, EM, token-F1
RQ3	Error and robustness analysis	Per-class recall, head-tail gap, OOV/unknown rates
RQ4	UC severity-focused experiments	Macro-F1, balanced accuracy, QWK, remission sensitivity/specificity
RQ5	Scenario-driven analysis	Scenario pass rate, high-risk error counts, clinician-facing clarity criteria
RQ6	Retrieval-augmented prototype analysis	Accuracy delta, evidence relevance, factual consistency checks

Table 1.4. Research Question-to-Evaluation Map

1.6.2 Research Hypotheses (Working)

The following working hypotheses guide this dissertation and will be tested in later chapters:

- H1: Zero-shot free generation with constrained or supervised decoders is predicted to perform better on core reliability metrics when using up-to-date GI datasets.
- H2: When aggregate accuracy is high, minority severe classes are supposed to continue to be the bottleneck.
- H3: An evidence-enhanced layer (conditioned) should enhance interpretability of high-risk questions without visual grounding.

These are regarded as tentative hypotheses to be checked in the future instead of being definite statements.

1.7 Scope, Assumptions, and Delimitations

1.7.1 In Scope

- Image-question answering of colonoscopy and GI endoscopy.
- Clinical axis of UC severity-oriented analysis.
- Multi-dataset benchmarking, based on persisted local artifacts.
- Generative answer tracks and closed set.
- Design of an evidence-aware extension pathway (RAG/PICO-oriented direction).

1.7.2 Out of Scope in This Dissertation Phase

- Prospective real-time deployment in clinical endoscopy units.
- Full regulatory and compliance validation.
- Multi-center randomized outcome trials.
- latency optimization of full video stream production grade.

1.7.3 Working Assumptions

- Annotations based on ground truth are considered as operational reference standards with known inter-observer variation.
- Persisted repository artifacts are the reproducible evidence base of the claims of this study.
- Evidence-aware generation is considered to be an additive and guarded layer.
- High-risk low-confidence deliverables need either abstinence or escalation.

1.7.4 PICO-Oriented Evidence Mapping (Operational Definition)

PICO is applied as a structured retrieval and response model of management-style queries in this research:

P (Population/Patient): UC patient setting, such as the present endoscopic severity and other pertinent clinical characteristics.

I (Intervention): the treatment or management option under consideration.

C (Comparison): alternative treatment strategy, standard care or no escalation.

O (Outcome): a clinically significant endpoint, i.e. relapse, mucosal healing, risk of hospitalization or adverse events.

A query pattern that will be used in Chapter 5 is:

Does infliximab or vedolizumab decrease relapse in patients with UC and Mayo 2 activity at 6 months?

In this workflow, MedVQA layer offers the visually knowledgeable seriousness signal, and the evidence layer maps the textual request onto PICO parts, finds pertinent investigations or guidelines, and provides a citation-connected response with confidence data and clearly articulated constraints.

1.7.5 Safety and Clinical Governance Position

This research considers the proposed system as a type of clinical decision support and not autonomous diagnosis. Its outputs are to be used as draft recommendations, which should be reviewed by the physicians, where there should be abstinence or escalation in uncertain or high-risk cases. That is why the uncertainty management, references to evidence and reporting of failures are considered as the main requirements of the evaluation design, whereas the regulated deployment is not the focus of this stage.

1.7.6 Operational Definitions of Reliability and Validity

Reliability in this dissertation is the consistency and reproducibility of a MedVQA system given the task, dataset and evaluation scenarios. The system is considered more reliable when it can provide consistent, reproducible, and clinically relevant results across evaluation scenarios, instead of only performing well on the overall benchmark. Accordingly, we measure reliability in terms of accuracy, macro-F1, balanced accuracy, class-wise recall, ordinal agreement, thresholded slices (clinical), output parse adherence, confidence intervals, paired tests (where possible) and reproducibility via artifacts.

Validity is the degree to which the evaluation supports the claims. Validity is not used as a general term in the dissertation. Rather, validity takes four forms.

Internal validity involves whether the research findings are likely to be caused by the tested model or pipeline, and not by data leakage, inconsistent training-test data splits, incomplete research artifacts, or uncontrolled preprocessing modifications.

External validity concerns whether results will hold up to broader datasets and fixed benchmark evaluation settings, such as different institutions, devices, image distributions, or clinician workflows.

Construct validity concerns whether the chosen evaluation metrics and procedures adequately measure the research constructs. In this work, the target constructs are reliability of GI MedVQA scores, agreement of UC severity scores, control of output format, visual grounding, evidence support, and usefulness to clinicians. For instance, quadratic-weighted kappa (QWK) and mean absolute error/root mean square error (MAE/RMSE) are used to evaluate Mayo severity because severity is ordinal, not categorical; token-F1 and BLEU-like metrics are only used as text surface diagnostics and not as proof of clinical correctness.

Conclusion validity concerns whether conclusions are justified by the evidence in terms of its size, stability, and strength. This is handled by multi-seed reporting, confidence intervals, paired tests (where possible), conservative reporting of missing metrics, and limiting claims to the data presented.

1.8 Scenario-Driven Framing

This research takes a situational approach to assessment. Aggregate benchmark metrics are still significant, yet do not encompass clinical utility. The scenarios are thus specified to capture real clinician questions and conditions of asymmetric risk.

1.8.1 Core Scenarios

Scenario ID	Clinical question style	Required output	Primary risk	Why it matters
S1	"What is the likely Mayo severity?"	Ordinal class + short rationale	Under-calling severe disease	Treatment escalation decisions
S2	"Is active bleeding visible?"	Binary answer + confidence	False negatives in high-risk frames	Urgent management context
S3	"How many polyps and where?"	Count + location-aware response	Missed findings or localization error	Procedural completeness and reporting
S4	"Summarize visible findings."	Multi-finding summary	Hallucinated findings	Interpretability and communication quality

S5	"Given severe findings, what evidence-backed options are relevant?"	Answer + evidence pointer	Weakly grounded management guidance	Decision-support extension

Table 1.5. Core Clinical Scenarios for Scenario-Driven Evaluation

All situations are meant to reveal a failure of a different nature. S1 and S2 have the risk-sensitive reliability first. S3 focuses on grounding and structured output fidelity. S4 checks the quality of controlled generation. S5 assesses the limit of visual interpretation and evidence-based reasoning.

1.8.2 Scenario Acceptance Logic

Question received

1. visual-grounded answer candidate

2. Checks confidence and consistency

3. if high-risk and low confidence: abstain/escalate

4. in case query (management-style): turn on retrieval evidence layer.

5. return clinician-facing response

1.8.3 Scenario-to-Metric Mapping

Scenario	**Priority metrics**
S1	Macro-F1, QWK, severe-class recall
S2	Sensitivity/recall, NPV, abstention behavior
S3	Count error (MAE/RMSE), location correctness slices
S4	Token-F1/ANLS + hallucination audit
S5	Answer correctness + evidence relevance/consistency checks

Table 1.6. Scenario-to-Metric Mapping

This mapping is used in later chapters to support a tiered acceptance model:

- Technical pass: minimum quantitative criteria are passed.
- Scenario pass: performance is acceptable for the scenario-specific risk profile.
- Workflow pass: output format supports clinician interpretation and action.

1.9 Conceptual Framework and System Flow

This section outlines the end-to-end architecture, design logic, and chapter dependencies followed throughout the dissertation. Three linked figures are used to summarize this structure. Figure 1.1 presents the clinician-facing MedVQA pipeline, Figure 1.2

shows how raw datasets are converted into evaluation evidence and decision-support readiness, and Figure 1.3 maps the research questions to the methods, metrics, and dissertation outputs. Together, these figures explain how the dissertation moves from clinical motivation to reproducible experimentation and finally to a guarded physician-support workflow.

1.9.1 End-to-End Conceptual Flow (Text Form)

Combination of Clinical Image(s) and Clinician Question.

1. intent parsing of questions (task type, entities, risk level)
2. extraction of visual features (global and localized cues)
3. cross-modal reasoning (closed-set or generative path)
4. answer candidate + confidence estimation
5. conditional retrieval layer for evidence-oriented queries
6. final response (answer, rationale, confidence, evidence pointer)

1.9.2 Figure 1.1: Clinician-Centric MedVQA Pipeline

Clinician-Centric MedVQA Pipeline for Colonoscopy.

Input frame(s)+question -> intent parsing + clinical entity extraction -> visual encoder + localization cues -> cross-modal reasoning -> confidence/abstain gate -> constrained or generative answer -> optional retrieval grounding -> clinician-facing response package

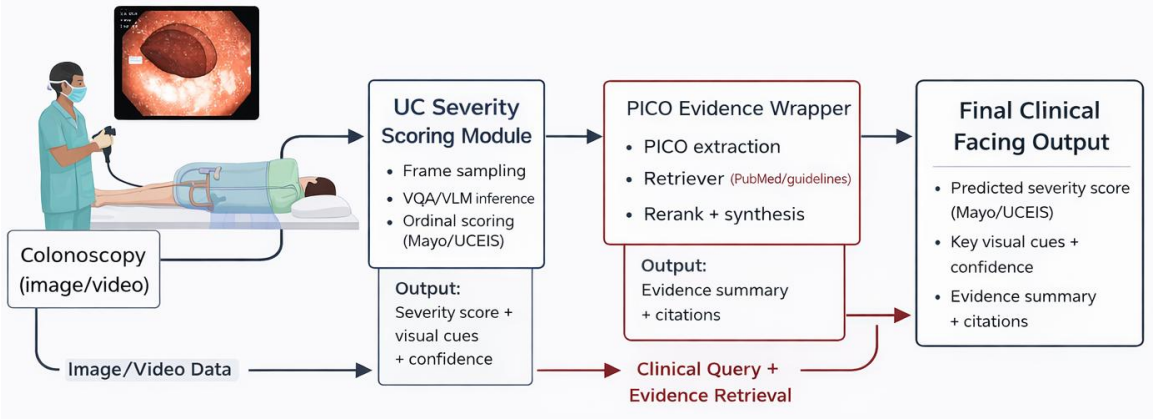

Figure 1.1. Conceptual Framework & System Flow

1.9.3 Data-to-Decision Funnel

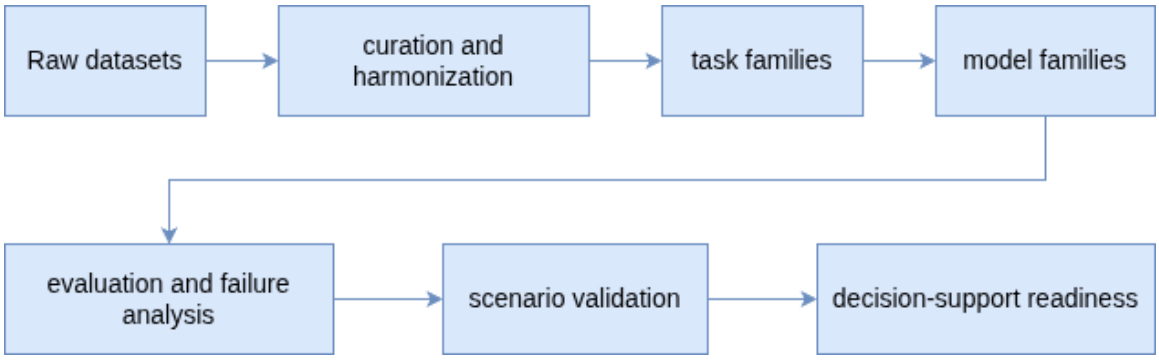

Figure 1.2. Data-to-Decision Funnel.

Figure 1.2 summarizes the dissertation's data-to-decision logic. The figure shows that the work does not move directly from raw datasets to clinical claims. Instead, raw GI datasets are first curated and harmonized, then organized into task families, evaluated across model families, examined through failure analysis, and finally interpreted through scenario-level validation before any decision-support claim is made.

1.9.4 Research Design Map

Figure 1.3 maps the research design used in the dissertation. It links the research questions introduced in Section 1.6 to the dataset and method layer, the metric layer, and the scenario acceptance layer. This map is included to show that each research question is connected to a specific form of evidence rather than being answered only through general discussion.

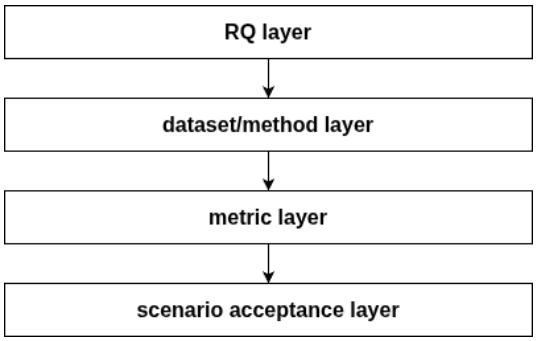

Figure 1.3. Research Questions to Methods to Metrics to Outputs.

Together, Figures 1.1, 1.2, and 1.3 provide the conceptual foundation for the remainder of the dissertation. Figure 1.1 explains the system flow, Figure 1.2 explains the evidence-building process, and Figure 1.3 explains how the research questions are operationalized through datasets, methods, metrics, and outputs.

1.9.5 Risk-Control Matrix

Risk	Failure example	Control strategy
Class imbalance	Severe UC under-detected	Cost-sensitive training, minority sampling, classwise monitoring
Lexical mismatch	Free text not mappable to clinical label space	Constrained decoding and lexical projection safeguards
Hallucination	Non-visible finding asserted	Visual-grounding checks, contradiction tests, abstain option
Overconfidence	High confidence on incorrect severe class	Calibration, thresholding, uncertainty-triggered escalation
Context gap	Image answer lacks decision context	Conditional retrieval-backed evidence layer

Table 1.7. Risk Controls for GI MedVQA System Design

1.10 Contributions of This Dissertation

There are five main contributions of this dissertation.

1. It creates a unified GI MedVQA evidence base across various datasets and task formulations with reproducible local artifacts, which enables one to compare performance in closed-label, generative and ordinal clinical environments.

2. It demonstrates by cross-dataset empirical analysis that constrained and supervised pipelines are more robust to naive zero-shot vision-language generation on current GI endoscopy tasks, especially in the case of class imbalance, answer-space mismatch, and format-sensitive evaluation.

3. It learns a controlled generative ulcerative colitis severity-scoring pipeline on Mayo 0-3 prediction on LIMUC and demonstrates, under frozen multi-seed internal analysis, that the LoRA-adapted generative mode1 lane is more accurate, has higher macro-F1, balanced accuracy, and quadratic weighted kappa, and has preserved perfect compliance to the parsing.

4. It presents a GenAI physician-query support wrapper based on PICO and an indication of retrieval, citation linkage, refusal behavior, and completion-audit traceability, and it shows reproducible base-level performance with definite claim limits.

5. It provides a reproducibility-first, claim-bounded colonoscopy clinician-facing multimodal AI framework demonstrating how model-level reliability, evidence-based interaction, and safety-focused output controls can be structured into a research-grade defensible decision support pipeline.

1.11 Chapter-by-Chapter Summary

Chapter 1 establishes clinical motivation, research problem, aims, objectives, research questions, scope limitations, conceptual framework, and contribution structure of the dissertation.

Chapter 2 discusses the visual question answering in medicine with special reference to gastrointestinal endoscopy and colonoscopy. It discusses families of datasets, technique families, assessment practices, and more recent advances, and it determines the gaps in translation that inspire this dissertation.

Chapter 3 introduces an empirical study with available VQA methods in datasets of GI-endoscopy based on persisted repository artifacts. It contrasts constrained and supervised pipelines and naive zero-shot generative methods and reveals the major reliability bottlenecks that have to be resolved prior to clinician-facing deployment can be discussed.

Chapter 4 builds up the suggested generative vision-language model of ulcerative colitis severity grading. Based on LIMUC as the main evidence base, it specifies a controlled Mayo 0-3 scoring task, sets a powerful supervised anchor, deploys the LoRA-based generative adaptation, and measures the ensuing system in a frozen multi-seed protocol with clear claim limits.

Chapter 5 builds upon the severity scoring on a model level and transfers it to physician-facing interaction by introducing a PICO-based GenAI wrapper. It assesses retrieval, citation-linked synthesis, refusal behavior, and auditability as constrained downstream elements that are constructed on the fixed severity evidence as developed in Chapter 4.

Chapter 6, wraps up the dissertation by summarizing the findings of Chapter 1 to 5 in answering the research questions, summarizing the final contributions, discussing the practical implications and limitations and giving directions on future research.

Chapter 2.

Surveying VQA Techniques in Medicine with Application to Colonoscopy

2.0 Chapter 2 Overview

Chapter 2 gives the scoping-review background of this dissertation and outlines the evidence landscape on which the empirical work in Chapter 3 is based. It starts by establishing the review design, which comprises the purpose, search window, type of sources, and inclusion and exclusion criteria. It then follows the evolution of MedVQA to constrained discriminative pipelines into transformer-based and generative multimodal models. Continuing on that wider path, the chapter then gets more specific to GI endoscopy with its key datasets, challenge tracks, and colonoscopy-specific task conditions, and is concerned with the reliability issues, including the presence of class imbalance, lexical variability in responses, vulnerability to clinical decision limits. It also contrasts the large families of techniques, such as discriminative, transformer-based, generative, explainable, and retrieval-aware techniques, with the scenario-level requirements of colonoscopy decision support. In order to enhance traceability of the thesis, the chapter takes into account both external literature and internal repository evidence and ends with a structured map of the key gaps. Such gaps are not considered to be abstract research opportunities, but they are discussed as tangible methodological imperatives guiding the design of Chapter 3, the choice of priority metrics and limits of the assertions presented in this dissertation.

Combined, this chapter constitutes a clear scoping framework, overview of methodological and benchmark development of MedVQA, mapping of GI-colonoscopy datasets and challenge ecosystems to task-relevant needs, comparing technique families with scenario-conscious clinical requirements, and generates gap-and-risk synthesis directly inspiring the experiments in Chapter 3.

The following chapter shifts synthesis of literature to empirical assessment of artifact, based on GI-endoscopy datasets.

2.1 Introduction

Visual Question Answering (VQA) has grown into a multimodal, clinically relevant research domain, out of a general computer vision benchmark task. The medical VQA (MedVQA) requires a model to interpret medical images, comprehend a natural-language query, and respond with a not only technically accurate but also clinically relevant answer. This imposes more demands on the system, such as more visual grounding, domain-sensitive language comprehension, ability to manage class imbalance, and safe behavior in the face of clinical risk, than general-domain VQA (Dong et al. 2025).

The subject of this dissertation is gastrointestinal (GI) endoscopy, especially cases of colonoscopy. This is a clinically significant, yet methodologically difficult field. Artifacts, fluctuation of light, blur, instrument occlusion and fine lesion morphology may be present in colonoscopy frames. Seemingly trivial questions in natural language, like Is there active inflammation? or What is the likely severity? might demand fine-grained visual processing and domain-specific interpretation.

This chapter has the aim of giving a rigorous survey that guides the design decisions in the subsequent chapters.

- maps the historical evolution of MedVQA model families.
- gathers data and benchmarks development, GI-specific focus.
- compares technique families with the requirements of colonoscopy scenarios.
- reviews evaluation practices and their limitations.
- identifies open gaps that directly motivate the methodology in Chapter 3, Chapter 4, and Chapter 5.

2.1.1 Definitions and Notation Used in This Chapter

Term / Metric	Working meaning in this dissertation
MedVQA	Image-question-answer modeling in medical domains
Clinical MedVQA	MedVQA configured for decision support with safety constraints
VLM / MLLM	Vision-language model / multimodal large language model
MES / UCEIS / BBPS	Mayo Endoscopic Subscore / Ulcerative Colitis Endoscopic Index of Severity / Boston Bowel Preparation Scale

EM / macro-F1	Exact match (string-level) / class-balanced F1 across labels
QWK	Quadratic weighted kappa for ordinal agreement (used in severity tasks)
ECE	Expected calibration error for probability-confidence alignment

Table 2.1. Working Definitions and Metric Notation

2.2 Scoping Review Design

2.2.1 Review Objective

The chapter follows a scoping-review methodology, but not a formal systematic-review protocol. The objective is to cover methodologically and support in practice thesis design as opposed to comprehensive bibliometric accounting. Sources were chosen to include:

- foundational MedVQA datasets and methods,
- modern multimodal large model directions,
- GI-specific benchmark and challenge resources,
- explainability and retrieval-grounding directions,
- and evaluation methodology relevant to clinical deployment.

2.2.2 Search Window and Source Types

The review window for this chapter spans 2018 to early 2026.

Primary source types included:

- peer-reviewed papers and data descriptors,
- official benchmark/challenge pages,
- official challenge overview papers and proceedings,
- and influential preprints where they define active benchmark directions.

Databases and discovery channels used in this scoping pass:

- PubMed / MEDLINE
- IEEE Xplore
- ACL Anthology
- arXiv
- challenge portals (ImageCLEF, MediaEval) and official dataset repositories

2.2.3 Query Themes

Theme	Example queries
General MedVQA foundations	medical visual question answering dataset, VQA-RAD, PathVQA, SLAKE
Modern model families	LLaVA-Med, Med-Flamingo, BLIP-2 medical VQA, medical LVLM benchmark
GI-specific resources	Kvasir-VQA, Kvasir-VQA-x1, ImageCLEF MEDVQA-GI, MediaEval Medico 2025
Evaluation and reliability	medical AI evaluation metrics, MedVQA robustness, hallucination medical VQA
Explainability and grounding	explainable medical VQA, multimodal explanation GI VQA, retrieval-augmented medical VQA

Table 2.2. Query Themes Used for the Scoping Review

2.2.4 Inclusion and Exclusion Logic

Type	Criteria
Inclusion	Primary technical work on MedVQA methods, datasets, benchmarks, or challenge design
Inclusion	GI/endoscopy resources directly usable for colonoscopy-oriented VQA analysis
Inclusion	Official benchmark/task pages and official challenge overview papers
Inclusion	Recent preprints with direct methodological or benchmark impact (explicitly marked as preprint where relevant)
Exclusion	Opinion pieces without technical evidence
Exclusion	Sources with no direct relation to visual-question answering in medicine
Exclusion	Tertiary summaries when primary sources were available

Table 2.3. Inclusion and Exclusion Criteria

2.2.5 Core Study Pool

This chapter synthesis is based on a hand-selected collection of historical and current sources within datasets, models, evaluation, and challenge design. The pool of sources is narrow enough to be relevant to the development of the field but is sufficiently broad to allow actionable methodological choices to be made to address the thesis.

To be transparent, a PRISMA-ScR-style accounting that was rebuilt based on the drafting log followed when triaging the source is also present in this chapter.

Stage	Count
Records identified across sources	236
Duplicates removed	54
Records screened (title/abstract)	182
Records excluded at screening	96
Full-text records assessed	86
Full-text records excluded	45
Included in chapter synthesis (external)	41

Table 2.4. Scoping Review Screening Summary

Exclusion reason	Count
Not MedVQA-specific (generic CV/NLP without QA task linkage)	16
Limited clinical/GI relevance for this thesis scope	11
Non-primary or tertiary source where primary source existed	8
Insufficient methodological detail for comparative synthesis	10

Table 2.5. Full-Text Exclusion Reasons in the Scoping Review

2.3 Evolution of VQA Methods: From General AI to Clinical MedVQA

The technical development of MedVQA follows, with some delay, the broader trajectory of VQA research in AI.

2.3.1 Stage A: Early Discriminative Pipelines

The original VQA systems usually used shallow fusion with CNN-based image encoders and question encoders based on RNN. The last task in these models was typically in the form of answer classification based on a predetermined set of vocabulary. This design was somewhat data efficient and operationally stable when the set of questions is constrained, but it had little flexibility with open-ended clinical answers.

2.3.2 Stage B: Attention and Cross-Modal Transformers

Multimodal networks based on transformers enhanced cross-modal alignment and reasoning capacity. ViLBERT and LXMERT are two-stream and cross-attention models that contributed to the development of a more robust pretraining paradigm of vision-language tasks (Lu et al. 2019; Tan and Bansal 2019). Transfer learning became

larger-scale with large-scale contrastive pretraining as in models like CLIP (Radford et al. 2021).

2.3.3 Stage C: Instruction-Tuned Generative Multimodal Models

More modern systems are coming to view MedVQA as a generative task as opposed to a fixed-label classification problem. Such a shift is visible in the BLIP-2-like adaptation (J. Li et al. 2023), LLaVA-like visual instruction tuning (Liu et al. 2023), and medical-domain adaptation variants like LLaVA-Med (C. Li et al. 2023) and Med-Flamingo (Moor et al. 2023). At this step, biomedical language pretraining backbones like BioBERT and BioGPT (Lee et al. 2019; Luo et al. 2022) are also pertinent since it enhances terminology fidelity and domain-oriented language control in downstream medical question-answering systems.

The advantage of this change is the possibility to facilitate more interaction and more explanation-based output. The primary risk, though, is that generated text produced fluent might be feebly attached to the image, lexically incompatible to benchmark labels, and even unsafe clinically unless it is carefully constrained.

2.3.4 Stage D: Reliability, Explainability, and Evidence Grounding

The present frontier goes further than answer generation to address requirements of clinical trust such as grounding, calibration, interpretability and evidence linkage. These encompass benchmark-quality research in reliability analysis (Hicks et al. 2022), multimodal in-context robustness (Hu et al. 2024), explainability frameworks (ImageCLEF 2024), and retrieval-based approaches in clinical VQA tasks (Luo et al. 2022).

Dataset / Benchmark	Year	Reported scope	Why it matters
VQA-RAD (Ben Abacha et al. 2019)	2018	Clinician-authored radiology QA	Early high-quality clinician-driven MedVQA benchmark
PathVQA (He et al. 2020)	2020	Pathology QA from textbook/digital resources	Extended MedVQA to histopathology
SLAKE (Liu et al. 2021)	2021	Bilingual, semantically labeled medical QA	Added richer semantic structure and multilinguality
PMC-VQA (Zhang et al. 2023)	2023	Large-scale visual instruction tuning corpus	Enabled modern generative MedVQA pipelines
OmniMedVQA (Hu et al. 2024)	2024	Large multi-dataset LVLM benchmark	Stress-tests generalization across modalities and anatomy
MedBookVQA (Yip et al. 2025)	2025	Textbook-derived multimodal benchmark	Structured benchmark for broad medical domains
MedFrameQA (Yu et al. 2025)	2025 (rev. 2026)	Multi-image reasoning benchmark	Closer to real clinical workflow than single-image QA

Table 2.6. Foundational and Recent General MedVQA Resources

Resource	Year	Reported scale	Role in this thesis
HyperKvasir (Borgli et al. 2020)	2020	110,079 images + 374 videos	Core GI visual foundation and class diversity
Kvasir-Capsule (Smedsrud et al. 2021)	2021	117 videos, 4.7M+ frames, 47k+ labeled frames	Robustness and broader GI variability context
LIMUC (Polat et al. 2022)	2022	11,276 UC images from 564 patients	UC severity and ordinal evaluation anchor
ImageCLEF MEDVQA-GI 2023 (Murtaza et al. 2023; ImageCLEF 2023)	2023	Multi-subtask GI VQA/VQG/VLQA benchmark	First major GI-specific VQA shared-task setup
Kvasir-VQA (Gautam et al. 2024; Simula Datasets n.d.)	2024	6,500 images, 58,849 QA pairs	Main GI text-image pair benchmark
ImageCLEF MEDVQA-GI 2024 (ImageCLEF 2024)	2024	Second-year challenge, synthesis-linked direction	Signaled task broadening around synthetic workflows
Kvasir-VQA-x1 (Gautam et al. 2025a; Simula n.d.)	2025	159,549 QA pairs with complexity levels and perturbations	Stronger reasoning and robustness benchmark
ImageCLEF MEDVQA 2025 (ImageCLEF 2025)	2025	Third-year challenge with synthetic GI integration	Benchmark evolution toward real-synthetic design
MediaEval Medico 2025 (MediaEval 2025; Gautam et al. 2025c)	2025	GI VQA + multimodal explanation subtask	Explicit shift toward explainable clinical interaction

Table 2.7. GI Endoscopy Dataset Ecosystem Relevant to This Thesis

In all these resources, the type of supervision and evaluation varies in a manner that directly affects the way in which model performance ought to be understood. Kvasir-VQA and Kvasir-VQA-x1 primarily offer paired image-question-answer supervision, with heterogeneous answer space and lexical shortcuts and yes/no priors capable of artificially inflating apparent performance (Karim and Uzuner 2025; Lee et al. 2019). In comparison, ImageCLEF and MediaEval tracks add challenge-level protocol constraints, and extended assessment bundles, which involve classification quality, overlap-based measures, and explanation-based evaluation (Li et al. 2025; Lin et al. 2023; Liu et al. 2021). These differences are considered methodological design constraints, as opposed to noise in the benchmarks, in this research.

2.4 Dataset and Benchmark Landscape

2.4.1 Challenge-Level Benchmark Progression

The GI challenge ecosystem has progressed at a high pace:

- 2023: proposed GI VQA/VQG/VLQA as a specific challenge design (Murtaza et al. 2023; ImageCLEF 2023).
- 2024: extended to synthetic-image-oriented workflows (ImageCLEF 2024).
- 2025: ImageCLEF continued the MEDVQA-GI direction, while MediaEval introduced GI VQA with multimodal explanation tracks (ImageCLEF 2025; MediaEval 2025; Gautam et al. 2025c).

This development is significant since it alters the state of the art. Comparable target does not merely focus on correctness of answers anymore, but also on the quality of explanation, visual alignment, and clinical applicability.

2.4.2 UC Severity Automation Literature Bridge

There is a significant pre-VQA literature on UC severity automation, which will require this background to place the current research appropriately. Representative studies are deep learning-based grading vs human review in JAMA Network Open (Stidham et al. 2019), CAD-based endoscopic activity scoring in Gastroenterology (Ozawa et al. 2020), prospective multicentre CAD-based inflammatory activity in Gastrointestinal Endoscopy (Yao et al. 2023), and more recent MES/UCEIS-oriented deep neural modeling in Journal of Crohn's and Colitis (Takenaka et al. 2023).

Study	**Input type**	**Output target**	**Main contribution**	**Remaining limitation for this thesis**
(Stidham et al. 2019)	Endoscopic images	Disease severity grade	Human-level comparative severity scoring analysis	Limited interactive QA and evidence-linked outputs
(Ozawa et al. 2020)	Colonoscopy images	Endoscopic disease activity	CAD feasibility for UC severity grading	Primarily fixed-score prediction interface
(Yao et al. 2023)	Multicentre colonoscopy data	Inflammatory activity	Prospective multicentre evaluation	Not structured around clinician question-answer workflows
(Takenaka et al. 2023)	Endoscopic images	MES/UCEIS predictions	Strong recent MES/UCEIS-oriented deep model validation	No retrieval-grounded management response layer

Table 2.8. Selected UC Severity Automation Studies Before MedVQA Framing

These studies demonstrate that it is possible to use AI to score UC, but do not specifically consider the MedVQA issue that is the focus of this dissertation: how to combine severity estimation reliability with controlled interactive reasoning and evidence

based response generation. This gap is further noted in the editorial discussions and variability analyses, which indicate the constant heterogeneity in scoring and the persistence of a translational barrier (Polat et al. 2022; Radford et al. 2021).

2.5 Survey of Technique Families

This section compares the main method families with particular attention to the requirements of colonoscopy VQA.

2.5.1 Rule-Based and Heuristic Logic

Rule based systems encode expert knowledge by hand written criteria, like color thresholds, morphology hints, or hand written scoring rules. They are mostly strong in being transparent and determining. Their primary weaknesses lie in frailty to visual variation, as well as to be extended to open-ended language tasks.

Rule-based elements can also be applied in the context of colonoscopy as supporting constraints, such as quality filtering or post-hoc verification, but are not complete MedVQA systems.

2.5.2 Discriminative CNN/RNN and Early Fusion Models

These pipelines code the image and the question individually and then recombine the two representations to make the closed-set prediction. They are also good foundations of constrained yes/no or categorical problems where the space of answers is fixed and deterministic products are desirable.

They have weaknesses in compositional reasoning and flexibility to generate longer explanatory answers.

2.5.3 Transformer-Based Multimodal Fusion

Transformer models enhanced contextual reasoning and image-question alignment by using cross-attention. Multimodal encoders can be pre-trained on a different task and then modified more sample-efficiently to clinical VQA as compared to similar systems that were trained on it directly.

This family tends to be very reliable on structured answer spaces in GI settings, particularly when question templates and answer ontologies are well-behaved.

2.5.4 Generative MLLM-Based MedVQA

Generative methods generate free-text responses and enable more interactive forms. Notable contributors to this change are instruction tuning and parameter efficient

adaptation. Examples of representative medical systems are LLaVA-Med (C. Li et al. 2023), Med-Flamingo (Moor et al. 2023), and instruction-tuned large-scale resources such as PMC-VQA (Zhang et al. 2023).

Benefits:

- richer language output,
- rationalization of responses,
- greater compatibility with conversationally oriented interfaces.

Risks:

- hallucination,
- weak visual grounding,
- instability in answer-format on closed-label tests,
- and clinically vague verbosity.

2.5.5 Explainable and Grounded MedVQA

Explainability has become a benchmark-quality requirement in GI challenges (Lin et al. 2023; Liu et al. 2021). The aim is not to answer but to justify and localize reasoning cues in a manner that can be audited by clinicians.

The most recent work has explicit multi-component explainability pipelines (ImageCLEF 2024), challenge systems that combine QA and explanation generation, as well as localization (Liu et al. 2023).

2.5.6 Retrieval-Augmented and Evidence-Aware Methods

Retrieval augmentation is increasingly explored to ground answers in relevant examples or external evidence. Recent shared-task evidence in medical VQA (MEDIQA-WV 2025 overview and system reports) shows that retrieval-aware pipelines can improve schema adherence and answer usefulness when retrieval quality is explicitly controlled (Lu et al. 2019; Luo et al. 2022).

For this thesis, retrieval is decomposed into three operational modes:

- **Text-evidence retrieval:** guidelines, review papers, and trial summaries for management-style questions.
- **Case-based retrieval:** nearest-neighbor retrieval of visually similar prior GI cases for comparability support.

- **Hybrid multimodal retrieval:** question intent + image cues used jointly to route evidence selection.

This dissection is a direct support of Chapter 5 PICO-oriented wrapper. The visual module initially responds to image-grounded queries (such as severity), then the evidence module translates management queries into PICO slots and retrieves citation-linked support and finally synthesizes responses.

In the case of colonoscopy MedVQA, retrieval should thus be considered as an extension layer not a substitute to strong visual grounding.

2.5.7 Comparative Family-Level Synthesis

Family	Typical output mode	Key strengths	Main risks	Best-fit scenarios
Rule-based / heuristics	Deterministic labels/rules	Transparent, predictable	Brittle under visual variation	Narrow quality-control checks
CNN/RNN discriminative	Closed-set labels	Stable on constrained tasks, efficient	Limited compositional reasoning	Binary/attribute tasks with fixed ontology
Transformer fusion	Closed-set or short text	Strong image-question alignment	Data and tuning sensitivity	Multi-category constrained QA
Generative MLLM	Free text	Rich interaction and explanation potential	Hallucination, lexical drift	Narrative support with safeguards
Explainable MedVQA	Answer + rationale/grounding	Improves trust and auditability	Evaluation standardization still maturing	Clinician-facing decision support
Retrieval-augmented	Answer + retrieved evidence	Better factual anchoring in some settings	Retrieval error propagates to answer	Evidence-linked higher-level queries

Table 2.9. Technique Family Comparison for Colonoscopy MedVQA

2.6 Evaluation Practices in the Literature

One of the key results of this survey is that not only model architecture but also evaluation practice is the bottleneck.

2.6.1 Classification Metrics vs Clinical Risk

Accuracy and macro-F1 are often reported in many MedVQA papers. These are not enough and needed by clinical decision support. Aggregate measures can mask failure in harsh classes in imbalanced environments. This consideration aligns with more general advice in medical AI, which focuses on matching evaluation measures with task risk as opposed to basing them on one summary measure (Nguyen et al. 2025).

2.6.2 Generative Metrics and Their Limits

BLEU, ROUGE, METEOR, CIDEr, ANLS, token-F1, and exact match are useful to measure textual overlap, but none alone is sufficient to prove clinical correctness. A solution can be linguistically close to the reference yet clinically incorrect.

2.6.3 Calibration, Uncertainty, and Significance

Uncertainty-aware behavior is also needed in clinical deployment. Most studies however do not describe calibration diagnostics including anticipated calibration error, and they also do not include significant testing. This complicates reported improvements, and undermines confidence in their translational relevance.

2.6.4 Challenge Metric Profiles

- ImageCLEF GI tracks include classification and segmentation-oriented metrics, such as accuracy, F1, MCC, and region-based metrics where relevant (Murtaza et al. 2023; ImageCLEF 2023; ImageCLEF 2024; ImageCLEF 2025).
- MediaEval Medico 2025 includes text-overlap metrics plus clinical relevance and explainability-oriented components (MediaEval 2025; Gautam et al. 2025c).

2.6.5 Recommended Metric Bundle for This Thesis

The approach to this research employs a multi-layer metric approach:

- closed-set performance: accuracy, macro-F1, MCC, balanced accuracy
- generative overlap (where applicable): BLEU/ROUGE/METEOR, token-F1/ANLS
- severity-aware evaluation: QWK, remission sensitivity/specificity for UC slices
- reliability diagnostics: confidence intervals, paired tests, calibration where available
- scenario-level checks: high-risk error counts and acceptance criteria.

Metric family	Example metrics	Useful for	Known blind spot
Closed-set classification	Accuracy, macro-F1, MCC, kappa	Deterministic QA evaluation	Can hide minority severe-class failures if reported alone
Generative overlap	BLEU, ROUGE-L, METEOR, ANLS, token-F1	Text quality comparison	Weak proxy for clinical correctness
Ordinal/severity	QWK, MAE/RMSE over ordinal labels	Severity grading behavior	Not reported consistently across papers

Calibration/uncertainty	ECE, reliability slices	Risk-aware deployment analysis	Still rare in MedVQA reporting
Statistical robustness	CIs, McNemar/paired tests	Claim stability and significance	Frequently missing in benchmark papers
Expert review / explainability	Clinician relevance ratings, grounding checks	Clinical usability and trust	Higher annotation effort, protocol variability

Table 2.10. Recommended Metric Bundle for This Dissertation

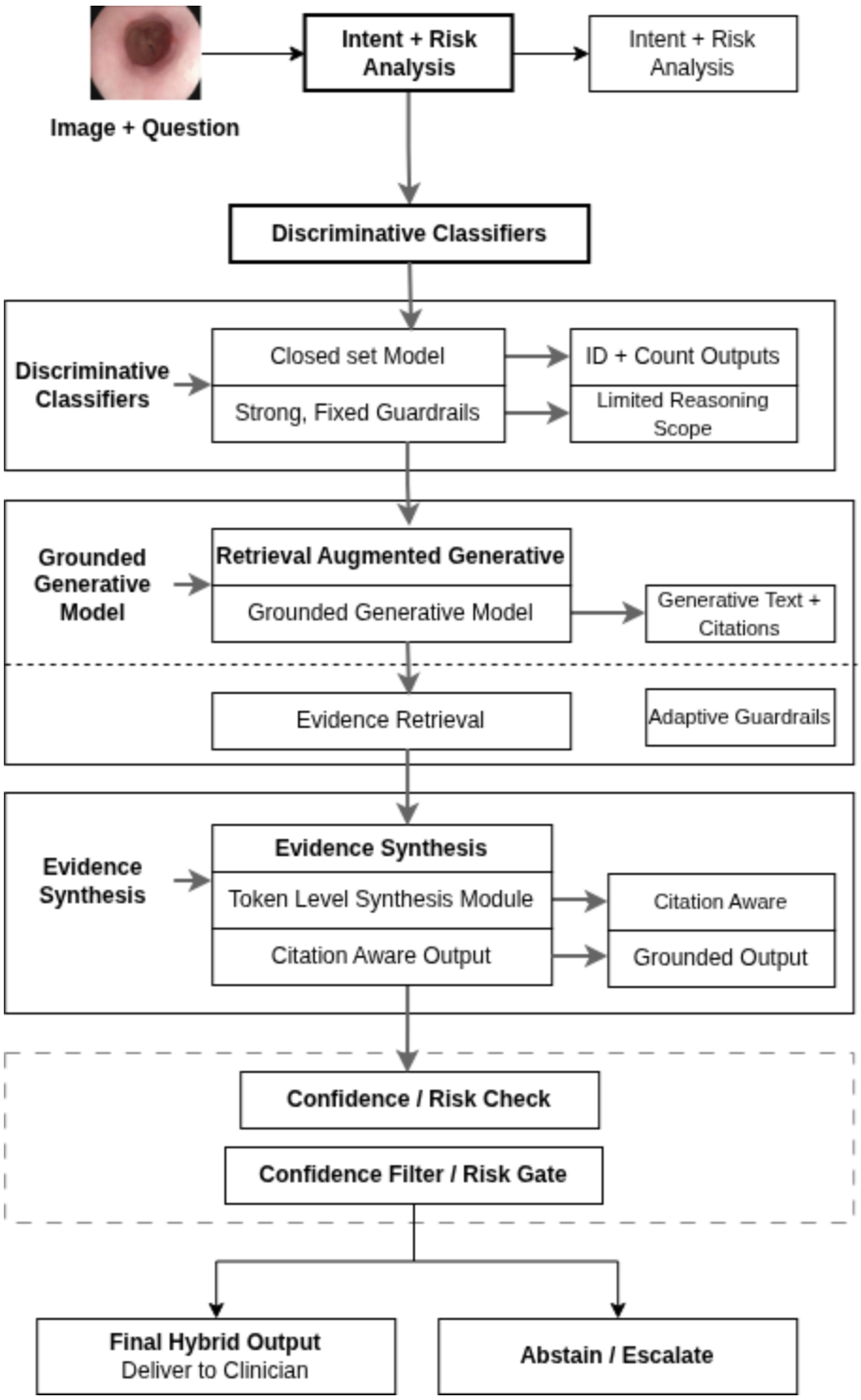

Figure 2.1: Hybrid Decision Support Routing Diagram

2.7 Scenario-Oriented Technique Suitability for Colonoscopy

A clinically useful survey should map model families to concrete clinical scenarios.

Colonoscopy scenario	Dominant question type	Preferred method profile	Reason
Finding presence/absence	Binary	Constrained discriminative or transformer classifier	Stable labels, high reliability requirements
Counting findings/instruments	Count + structured category	Transformer with constrained decoding	Better visual alignment than shallow fusion
Lesion location description	Spatial text/category	Transformer + localization-aware head	Requires spatial grounding consistency
UC severity grading	Ordinal category	Fine-tuned visual backbone + constrained QA interface	Ordinal robustness and severe-class control
Narrative explanation for clinician	Open text	Generative MLLM with grounding checks	Better usability, but must be constrained
Evidence-linked management query	Open text + evidence	Conditional retrieval-augmented generation	Useful only when core visual answer is reliable

Table 2.11. Scenario-Oriented Technique Suitability for Colonoscopy MedVQA

2.8 2024-2026 Trends in MedVQA Research

The recent literature shows that there are some significant changes.

2.8.1 From Single-Image QA to Multi-Image Reasoning

Multi-image clinical reasoning is challenging to current MLLMs and reveals a divide between benchmark fluency and longitudinal diagnostic reasoning, with MedFrameQA representing this multi-image clinical reasoning direction (Yu et al. 2025).

2.8.2 From Benchmark Accuracy to Reliability Stress Testing

The ProbMed style probing reveals that brittle behavior can be concealed by high benchmark scores when perturbed in a controlled manner (Hicks et al. 2022). SMMILE also emphasizes the weakness of multimodal in-context learning (Hu et al. 2024).

2.8.3 Data-Centric Expansion and Synthetic Pipelines

Semi-automated or automated pipelines are becoming increasingly popular to scale MedVQA data by using new resources. This is observed in MedVLSynther-style generator-verifier algorithms (Huang et al. 2025) and challenge tracks that specifically use synthetic data (Lin et al. 2021).

2.8.4 Scaling Unified Medical VLMs

OmniV-Med and other research are intended to assist in integrating multimodal medical knowledge across 2D, 3D, and video environments (Jiang et al. 2025). These models have potential in terms of general ability, but whether they can provide clinical robustness with consistency in terms of special GI question answering is an open question.

2.8.5 Explainability as a First-Class Objective

GI challenge design (2025) introduced quality of explanation and clinical relevance as explicit evaluation targets, instead of optional forms of analysis (Lin et al. 2023; Liu et al. 2021). This indicates an expedient move towards trust standards that are relevant to clinician facing environments.

2.9 Evidence Triangulation with Repository Results

To avoid relying on a purely narrative survey, this section triangulates findings from the literature with locally persisted artifacts.

Dataset/task (local report)	Representative local signal	Method implication
ImageCLEF MEDVQA-GI 2023	ViLT fine-tune val accuracy 0.9089, macro-F1 0.5823; zero-shot Qwen raw near-zero accuracy	Closed-set tuned models remain stronger than raw zero-shot generation in this setting
HyperKvasir 23-class	ResNet50 supervised outperforms saved zero-shot generative baseline by large margin	Robust supervised visual encoders remain critical for GI grounding
Kvasir-VQA	ResNet+GRU yes/no subset reaches high reliability; free generation runs can collapse to unknown outputs	Constrained answer spaces remain highly effective for deterministic clinical subquestions
Kvasir-VQA-x1	LoRA improves token-level generative metrics; exact-match remains challenging	Reasoning-rich QA increases complexity; output normalization and grounding are central
LIMUC severity	Fine-tuned ResNet50 leads macro-F1 and QWK; zero-shot VLM underperforms severe-class reliability	UC severity tasks favor domain-tuned supervised pipelines with ordinal-aware evaluation

Table 2.12. Repository Evidence Snapshot and Method Implications

2.9.2 Interpretation

The local results align with broader literature:

- constrained/fine-tuned pipelines are still the reliability baseline in GI tasks;
- generative models provide flexibility but require strict controls;
- evaluation must emphasize classwise and severity-aware behavior, not only aggregate scores.

2.10 Key Gaps and Open Problems

2.10.1 Data and Annotation Gaps

- Severe-class and rare-finding imbalance remains substantial.
- Cross-center, cross-device robustness evidence is still limited for GI VQA.
- Multi-image/video QA is growing but still comparatively immature.

2.10.2 Model and Grounding Gaps

- Open-ended outputs often outpace grounding reliability.
- General-domain zero-shot transfer remains unstable for specialized GI semantics.
- Explainability outputs are not yet standardized across benchmarks.

2.10.3 Evaluation and Reproducibility Gaps

- Metric heterogeneity hinders fair cross-study comparison.
- Many studies still lack calibration and statistical confidence reporting.
- Clinical utility is often inferred indirectly from generic NLP overlap metrics.

2.10.4 Translational Gaps

- Few studies evaluate full clinician-in-the-loop workflows.
- Escalation policies for uncertain/high-risk outputs are seldom formalized.
- Evidence-aware QA in GI remains promising but early-stage.

2.10.5 Threats to Validity (for This Survey and Downstream Design)

The principal validity threats that are addressed in this chapter are fourfold: (1) shift in the datasets among devices, centers and acquisition protocols (2) inconsistencies in the annotation of endoscopic severity labels (3) question ambiguity and the impact of answer-space normalization in generative systems and (4) mismatch in the metrics where the presence of linguistic overlap is not necessarily a sign of clinical accuracy.

The latter threats are extended to the empirical design of the subsequent chapters by split-aware testing, class-based testing, and scenario-level testing of acceptance.

Observed gap	Mitigation in this dissertation
Long-tail severe-class weakness	Class-aware and severity-aware evaluation slices (Chapter 3, Chapter 4)
Zero-shot instability in GI	Strong supervised and constrained baselines before open generation extensions
Metric mismatch with clinical risk	Multi-layer evaluation protocol with ordinal and imbalance-aware metrics
Explainability without standard criteria	Scenario-specific explanation requirements and clinician-oriented quality checks
Weak evidence linkage	Conditional retrieval-augmented extension with guardrails

Table 2.13. Gap-to-Thesis Mitigation Map

2.11 Chapter Summary and Transition

The chapter has provided a review of the MedVQA techniques in a colonoscopy-focused approach and has harmonized the results of the external literature and local repository artifacts.

A number of key findings can be made on the basis of this review. To begin with, MedVQA has developed a shift to fixed-label discriminative methods to generative multimodal systems that place a more significant focus on explainability. Second, GI benchmark resources have evolved at a pace since 2023, and more realistic evaluation is becoming feasible. Third, constrained and domain-tuned approaches are the best at providing reliable core behavior at present in high-risk colonoscopy cases. Fourth, generative and retrieval-enhanced methods are beneficial extensions, yet only when they are developed on the basis of strong visual grounding and tested under stringent conditions. Fifth, the critical concept of translational relevance requires scenario-driven evaluation that is consistent with clinical use.

These findings directly jump into Chapter 3, which takes the literature synthesis to the empirical exploration of the existing model families with the help of the datasets and persisted outputs that can be found in this repository.

Chapter 3.

Investigating Existing VQA Techniques Across GI-Endoscopy Datasets

3.1 Chapter Overview and Evaluation Goal

Chapter 2 surveyed key families of medical VQA techniques and found a critical translational divide in GI endoscopy: good benchmark performance is not always clinically reliable behavior, especially with class imbalance, domain shift and high-consequence failure modes (Simula n.d.; Smedsrud et al. 2021). Chapter 3 attempts to seal that gap in an empirical manner, through auditing persisted experimental artifacts stored in this thesis repository and summarizing the behavior of model families according to the dataset constraints that currently exist with colonoscopy-oriented MedVQA.

This chapter is intended to be the empirical turning point of the dissertation. The overall question is not only what model would score higher on aggregate scores, but what model behaviors are reliable enough to be thought of as building blocks to a set of clinician-facing decision support. To provide the answer to that question, the reliability of heterogeneous task regimes that take place in the GI MedVQA are evaluated: (i) closed-label prediction, (ii) open-ended generation, and (iii) ordinal severity assessment.

Chapter 3 overview (reader preview)

Across the datasets in this repository, the chapter provides a reproducible comparison of:

- constrained or supervised pipelines vs. raw zero-shot generative outputs,
- dataset and question-family-specific failure modes (imbalance, format drift, mapping fragility) and
- severity-oriented reliability under ordinal and clinically meaningful slices (e.g., remission vs. non-remission).

Notably, Chapter 3 is reproducibility-first: the results are reported only in case there are persisted artifacts. In cases where configuration or status files alone exist (such as scenario YAML definitions), the chapter reports only configuration, and does not purport to performance at the scenario level.

Practically, the goal is to establish, using traceable local evidence, the behavior of existing model families under:

- closed-set and binary clinical question answering,
- open-ended generative answering,
- severity-oriented ordinal assessment,
- class-imbalance stress conditions, and
- scenario-level stress-test protocol definitions for clinically relevant failure probing.

Chapter 3 is artifact-driven, unlike Chapter 2 (literature synthesis): all results are assembled from the repository artifacts listed in Appendix A rather than from new training or undocumented reruns. In the case where the prediction involved in a metric is not persisted the metric will not be reported.

Within the dissertation, this chapter serves three main functions:

1. It establishes a reproducible baseline of current GI MedVQA behavior before the introduction of new methods.

2. It identifies practical reliability bottlenecks, including sensitivity to class imbalance, instability across answer formats, fragility in lexical or out-of-vocabulary (OOV) mapping, and failures at clinically important thresholds.

3. It offers an empirically justifiable transition to Chapter 4, where controlled generative methods are formulated to generate clinically motivated UC severity rating, and subsequently extended to evidence-conscious responding.

Formally, Section 3.1 establishes evidence boundaries and research coverage, Section 3.2 formalizes datasets and model regimes, Section 3.3 defines evaluation metrics and caveats, Section 3.4 gives dataset-specific results, and Section 3.5 generalizes across dataset results to thesis-level conclusions.

3.1.1 Evidence Boundary and Reproducibility Scope

Table 3.1 maps each Chapter 3 evidence block to its role in the evaluation argument. The table shows what each dataset or artifact contributes to the reliability analysis, while retaining repository paths as source anchors for reproducibility.

Dataset / analysis block	Role in Chapter 3	Evidence contributed	Reliability issue examined	Primary outputs used later in the chapter	Reproducibility source
HyperKvasir	Visual-grounding baseline for GI imaging	23-class GI image classification results across supervised, frozen-feature, and zero-shot model settings	Long-tail class imbalance and whether aggregate accuracy hides weak minority-class behavior	Accuracy, balanced accuracy, macro-F1, MCC, kappa, and class-level interpretation	Prototyping_reformat/DatasetAnalysis/HyperKvasir/HyperKvasir.md; Prototyping_reformat/DatasetAnalysis/HyperKvasir/**/out
ImageCLEF MEDVQA-GI 2023	Closed-label GI VQA benchmark	Validation results for fine-tuned VQA and zero-shot generative variants under constrained answer IDs	Reliability gap between label-constrained VQA and raw or projected zero-shot generation	Overall validation metrics, question-family aggregates, and raw-to-projected diagnostic gains	Prototyping_reformat/DatasetAnalysis/ImageCLEF_MEDVQA_GI_2023/ImageCLEF_MEDVQA_GI_2023.md; Prototyping_reformat/DatasetAnalysis/ImageCLEF_MEDVQA_GI_2023/**/results
Kvasir-VQA	GI VQA subset analysis	Yes/no and attribute-style subset results from paired image-question-answer data	Sensitivity to question type, answer-space restriction, and majority-family effects	Yes/no results, attribute subset results, and source/question/answer-type profiling figures	Prototyping_reformat/DatasetAnalysis/Kvasir_VQA/Kvasir_VQA.md; Prototyping_reformat/DatasetAnalysis/Kvasir_VQA/**/out
Kvasir-VQA-x1	Large-scale generative VQA stress test	Free-text generation and mapped closed-set diagnostics for a larger GI VQA setting	Format instability, OOV behavior, lexical overlap limits, and mapped-accuracy inflation	Token-level generative metrics, mapped/baseline diagnostics, OOV analysis, and complexity-level summaries	Prototyping_reformat/DatasetAnalysis/Kvasir_VQA_x1/Kvasir_VQA_x1.md; Prototyping_reformat/DatasetAnalysis/Kvasir_VQA_x1/**/results
LIMUC	Flagship UC severity reliability axis	Mayo 0-3 severity classification results across frozen-feature, supervised, and zero-shot settings	Ordinal reliability, clinically meaningful threshold behavior, and minority severity-class bottlenecks	Overall test metrics, remission-vs-active slice metrics, and per-class best-model summary	Prototyping_reformat/DatasetAnalysis/LIMUC/LIMUC.md; Prototyping_reformat/DatasetAnalysis/LIMUC/**/out
Kvasir-SEG supporting analysis	Supporting morphology and localization context	Mask and morphology-oriented dataset statistics used to motivate grounding-aware evaluation	Whether visual grounding and spatial/morphological structure should be con-	Supporting dataset statistics and morphology/localization context	Prototyping_reformat/DatasetAnalysis/Kvasir_SEG/Kvasir_SEG.md; Prototyping_reformat/DatasetAnalysis/Kvasir_SEG/0_dataset_prep/**

			sidered alongside answer accuracy		
Scenario protocol configuration	Configuration evidence for stress-test design	Scenario definitions for clinically relevant failure probing	Protocol design for micro-scenario stress tests rather than evaluated model performance	Scenario categories and evaluation design constraints; not treated as result evidence	Prototyping_reformat/DatasetAnalysis/Kvasir_VQA/evaluation_comparison/scenarios/scenarios.yaml; Prototyping_reformat/DatasetAnalysis/Kvasir_VQA/evaluation_comparison/scenarios/
Legacy UC generative runtime snapshot	Reformatted runtime-status evidence	Historical open-ended UC response artifacts retained for traceability	Availability and status of older generative outputs, without treating incomplete outputs as new benchmark results	Runtime-status summaries and artifact availability checks	Prototyping_reformat/DatasetAnalysis/Kvasir_VQA/evaluation_comparison/phase3_results/summary_uc_phase3.csv; Prototyping_reformat/DatasetAnalysis/Kvasir_VQA/evaluation_comparison/phase3_results/*.csv

Table 3.1. Evidence Map and Reproducibility Sources for Chapter 3

Reproducibility rules used throughout Chapter 3

- Undocumented reruns: the chapter only reports what is in persisted artifacts and their report files.
- Report-first compilation: all subsections of the dataset in Section 3.4 are compiled using the dataset report file (named as mydata.md) first, and followed by any visualizations or summary outputs that are persisted as indicated by that report.
- No artificial comparisons: in cases where a model has prediction or labeling or scoring outputs that are not present, the chapter does not assume or extrapolate results.
- Configuration-only evidence is reported as configuration: scenario YAML files are not reported as evaluated results, but as protocol definitions.

- Local-report cardinalities: the sizes of datasets presented in this chapter are the counts reported in the local report artifacts and might not match the official totals of dataset descriptors because of subset selection, filtering, or split definitions.

3.1.2 Research Questions Addressed in This Chapter

Chapter 3 presents empirical data, which will be important to the subsequent research questions established in Chapter 1:

- **RQ2 (comparative reliability):** Do constrained/discriminative pipelines remain more reliable than raw zero-shot generative outputs for current GI MedVQA tasks?

- **RQ3 (failure modes):** What are the most common failure modes across datasets and task style (e.g., imbalance collapse, lexical drift, mapping brittle, question-family brittle)?

- **RQ4 (severity robustness):** How reliably can models support UC severity-oriented question answering under ordinal structure and severe-class imbalance?

- **RQ5 (clinical output format)** – protocol support only: Chapter 3 provides scenario protocol definitions to condition subsequent clinician-facing evaluation logic, but it does not provide scenario outcome measures unless scored predictions are stored.

In order to maintain the boundary of evidence, when scenario configuration alone exists, we defer scenario-level conclusions.

Chapter 2 has defined the technique landscape (discriminative fusion models, transformer-based multimodal encoders, generative multimodal models, and retrieval-augmented directions) and highlighted that evaluation should be in line with clinical risk (Smedsrud et al. 2021; Stidham et al. 2019; Takenaka et al. 2023). The operationalization of that principle in chapter 3 is based on the assessment of the model families in terms of the datasets and the artifact outputs provided in this repository. The chapter thus is the empirical reality test of the thesis. It determines what is already reliable given GI-specific constraints and the failure modes that should be resolved prior to controlled generative extensions and evidence-aware extensions being presented in the subsequent chapters.

3.2 Experimental Scenarios and Data Regimes

Section 3.2 then declares the empirical scope, which are the datasets under consideration, the types of tasks represented by the datasets, and the families of models under comparison. It is not an attempt to generalize reliability patterns in all GI environments, but to describe reliability patterns in the evidence of artifacts accessible and in different task regimes (closed-label, generative, ordinal).

3.2.1 Dataset-Task Matrix

Table 3.2 is a summary of the datasets that were used in Chapter 3 and the main purpose of each dataset in the assessment. The cardinality values are taken from the local report files and repository artifacts listed in Appendix A.

Dataset	Cardinality in local report	Core outputs	Main evaluation axis	Clinical relevance
HyperKvasir	10,662 images, 23 classes	class labels	multiclass reliability under imbalance	broad GI visual grounding
ImageCLEF MEDVQA-GI 2023	36,683 QA rows (29,351 train, 7,332 val)	label IDs per question	per-question closed-label VQA reliability	benchmarked GI QA consistency
Kvasir-VQA	58,849 QA rows, 6,500 images	yes/no, attributes, free text	subset reliability and format stability	colonoscopy QA behavior
Kvasir-VQA-x1	159,549 QA rows, 6,449 images	free-text answers + mapped labels	generative fidelity, complexity effects	robust MedVQA reasoning stress
LIMUC	11,276 images, Mayo 0-3	ordinal severity class	macro-F1, QWK, remission slice	UC treatment-aligned severity
Kvasir-SEG	1,000 image-mask pairs	mask morphology stats	coverage/shape support metrics	future localization grounding

Table 3.2. Dataset and Task Matrix for Chapter 3 Experiments

Why these datasets are included

- **Kvasir-VQA** offers the most direct VQA evidence that is colonoscopy-oriented and is applied to test subset reliability and answer-format stability in common question families (Safwan et al. 2025; Gautam et al. 2024).
- **ImageCLEF MEDVQA-GI 2023** offers a standardized closed-label GI VQA environment, in which fine-tuned VQA models are compared to zero-shot generative outputs subject to label constraints (Simula Datasets n.d.; Hu et al. 2021).
- **Kvasir-VQA-x1** is a large-scale generative stress test, which is used to examine overlap measures, complexity, and label-mapping vulnerability in free-text responses (Gautam et al. 2025a; Simula n.d.; Appendix A, Artifact A5).

- **LIMUC** supports the clinical flagship of the thesis: UC severity, in which ordinal structure and clinically meaningful slices (e.g., remission vs non-remission) play a crucial role in interpretation (Polat et al. 2022; Appendix A, Artifact A3).

- **HyperKvasir** provides broad GI visual grounding with strong long-tail imbalance, allowing analysis of head–tail recall asymmetry that can be hidden by aggregate accuracy (Rieff et al. 2025; Zhang et al. 2023).

- **Kvasir-SEG** is included as supporting evidence for morphology/localization-oriented statistics that later motivate grounding-aware directions; it is not treated as a primary VQA benchmark in this chapter (Appendix A, Artifact A6).

3.2.2 Model Families Compared

The comparisons of models in Chapter 3 are based on what exists as persisted artifacts, without a complete survey of all possible architectures. The idea is to compare representative families that cut across the key design options of clinical reliability: supervised classification, constrained VQA, and open-ended generation.

Family	Example persisted models	Typical answer mode
Supervised CNN/ViT classifiers	resnet50_supervised, vit_supervised, finetune_resnet50	closed label
Frozen encoder + shallow classifier	vit_frozen_logreg, clip_linear_baseline, resnet50_frozen_logreg	closed label
Classical multimodal fusion	resnet_gru_m1_*, vit_bertlite_m2_*	closed set
Transformer VQA fine-tuned	vilt_finetune	closed label per question
Zero-shot VLM/MLLM	qwen2_5_vl_zeroshot, medgemma_zeroshot, blip2_zero_shot	free text (optionally projected)
Parameter-efficient adaptation	medgemma_lora_original, qwen2_5_vl_lora_finetune (logs persisted)	free text

Table 3.3. Representative Model Families in Persisted Artifacts

Interpretation guide

- **Closed-label families** (supervised or shallow classifiers) have a baseline of reliability in which the output space is regulated and its evaluation is deterministic.

- **Classical fusion and transformer VQA models** represent structured MedVQA pipelines designed to combine image and question representations under a constrained output space.

- **Zero-shot VLM/MLLM models** represent open-ended generation, which may be more flexible but can be unstable in output format and lexical grounding.
- **Parameter-efficient adaptation** captures whether lightweight fine-tuning improves generative fidelity without requiring full retraining.

Section 3.3 defines the metric bundles used to compare these families and clarifies how raw vs projected scoring is treated when free text must be mapped to a constrained answer space.

3.2.3 Data Profile Figures (Kvasir-VQA)

The chapter provides Kvasir-VQA profiling figures before results are presented, as they describe some behaviors that were later on in Section 3.4, including yes/no question dominance, and skewness in templates. These distributional characteristics can boost superficial performance with majority question types and conceal failure with less frequent but clinically significant categories (Gautam et al. 2024; Appendix A, Artifact A4).

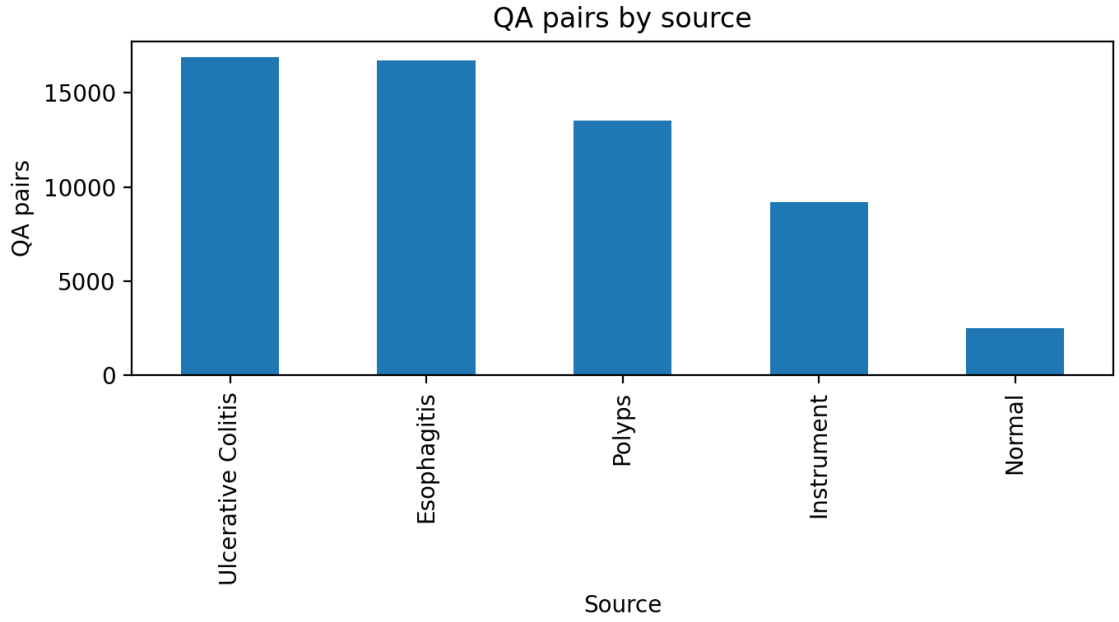

Figure 3.1. QA count by source domain

Figure 3.1 summarizes how QA pairs are distributed across source domains in the Kvasir-VQA preparation artifacts (Appendix A, Artifact A4). Source-domain imbalance implies that model performance may reflect dominant sources disproportionately, and it motivates later per-family and per-task reporting rather than single aggregate scores.

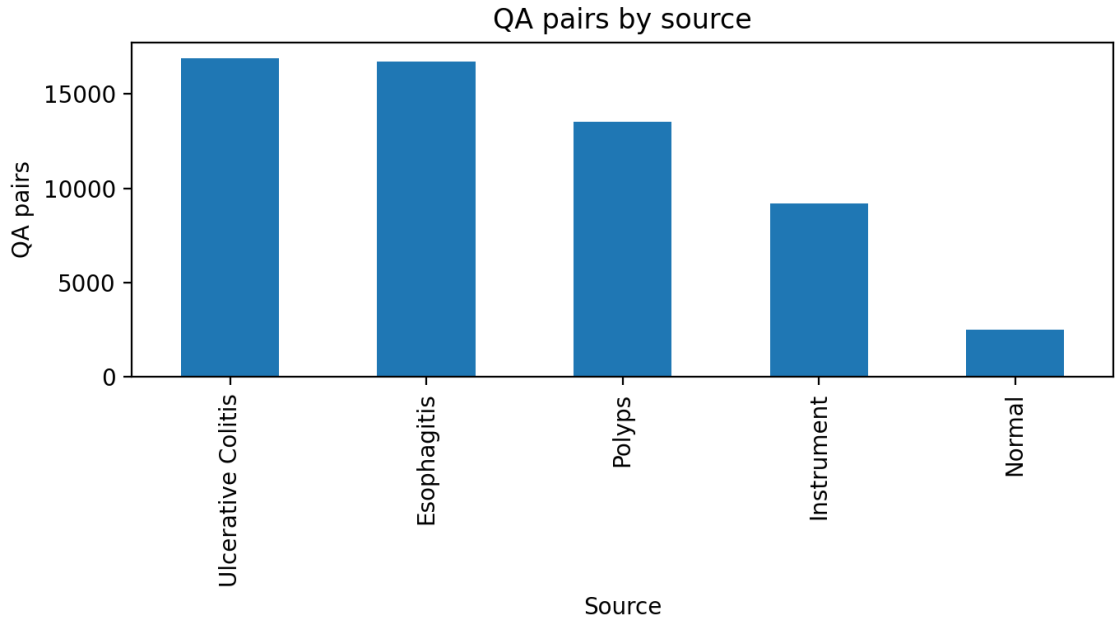

Figure 3.2. QA count by type distribution

The distribution shows that yes/no questions form a large share of Kvasir-VQA (Appendix A, Artifact A4). This creates a majority-family regime where models can perform strongly by exploiting priors, and it motivates the use of balanced metrics and subset-level reporting to avoid overstating reliability.

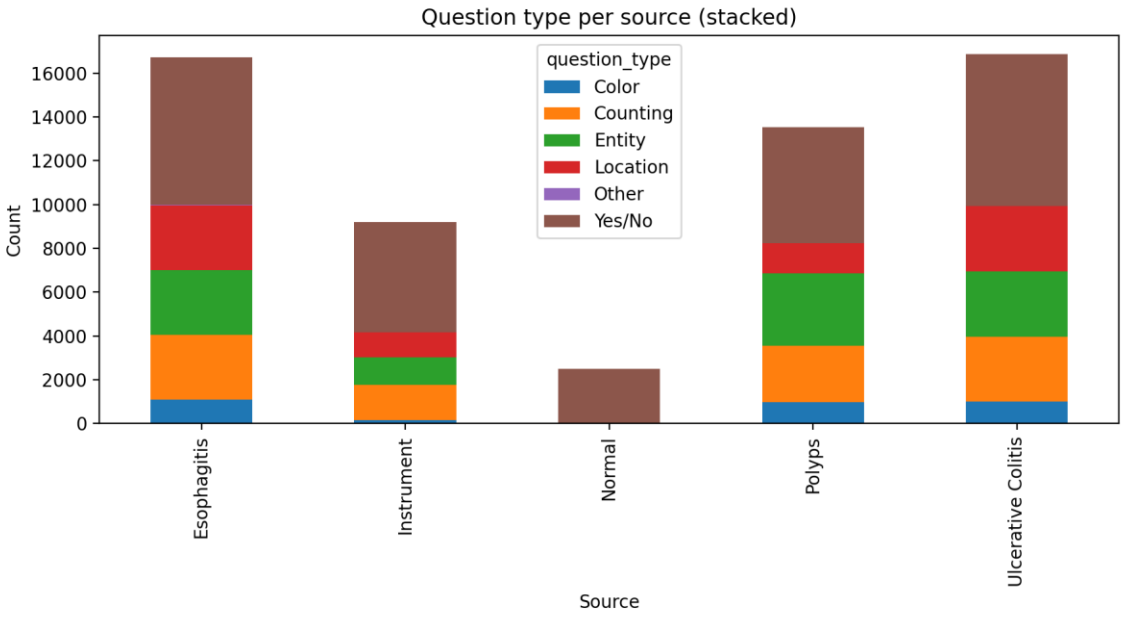

Figure 3.3. Question type by source

Question-type composition varies by source domain, indicating that the dataset is not homogeneous across origins (Appendix A, Artifact A4). This supports the chapter's emphasis on failure-mode analysis and warns against interpreting a single "overall" score as uniformly applicable across question families.

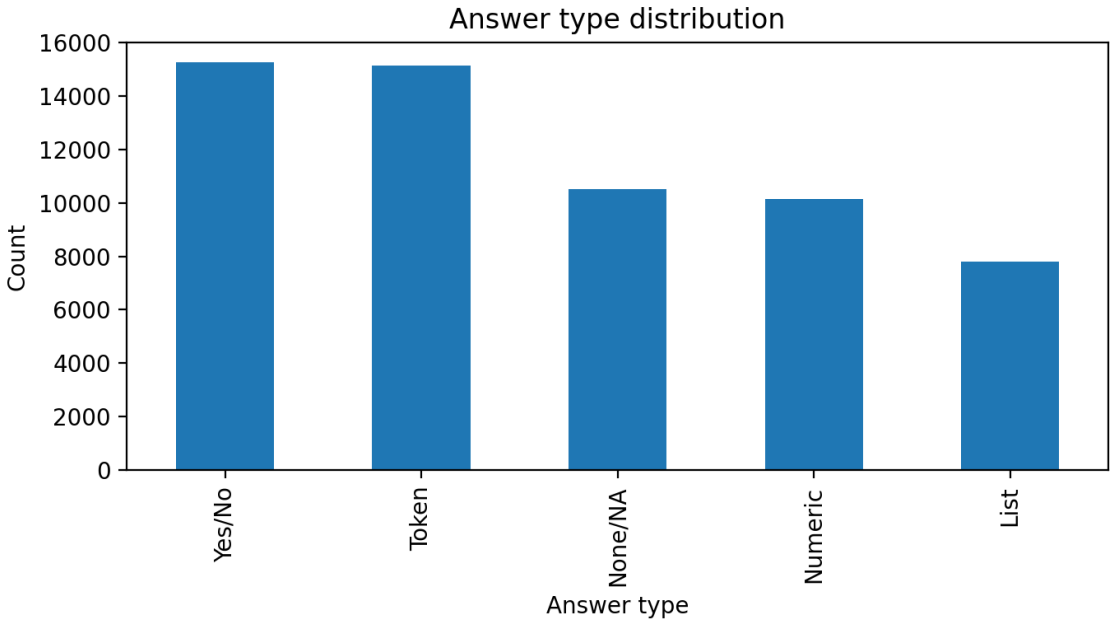

Figure 3.4. Answer type distribution

The answer-type distribution underscores the need to match evaluation to task style: exact-match measures are inherently suited to constrained outputs, but may be excessively severe for free-text generation, where token-overlap diagnostics can be more informative (Appendix A, Artifact A4). This inspires the stratified metric design given in Section 3.3.

Signal	**Value**
Total QA rows	58,849
Unique images	6,500
Mean QA rows per image	9.05
Yes/No questions	26,515 (45.06%)
Entity questions	10,528 (17.89%)
Counting questions	10,118 (17.19%)
Location questions	8,424 (14.31%)

Table 3.4. Kvasir-VQA Distribution Snapshot

The skew of the question family is large in favor of two design decisions made over the entire Chapter 3: (i) report class-balanced or family-aware measures where feasible, and (ii) be careful when interpreting aggregate scores, since good performance on

a dominant family (e.g., yes/no) does not imply good performance on count, location, or higher-risk categories..

3.2.4 Scenario Micro-Benchmark Definition

In order to bridge benchmark evaluation to real-world clinical vignettes, a micro-scenario protocol definition is contained in the repository.

The protocol outlines three scenario templates that are focused:

1. **S1**: active bleeding binary detection,
2. **S2**: instrument type and polyp count,
3. **S3**: Paris morphology closed set.

These are deliberately small scenarios that are not intended to be used as substitutes of statistical benchmarks; rather, they are considered as stress vignettes. The reformatted evidence tree only persists scenario configuration; the prediction of scored scenarios is not persisted. Thus, Chapter 3 records protocol contexts only of scenario definitions but not scenario outcome measures in the absence of scored artifacts.

3.2.5 Alignment with Prior Dissertation Problem Settings

Chapter 3 values and dataset and task selections is tailored to align with typical GI-endoscopy AI issues and concentrate on MedVQA-specific reliability:

- **Severity and treatment relevance** are reflected by LIMUC, in which ordinal behavior and clinically interpretable slices are required for credible UC severity assessment (Polat et al. 2022; Stidham et al. 2019; Ozawa et al. 2020; Yao et al. 2023; Takenaka et al. 2023; Appendix A, Artifact A3).

- **Robustness under heterogeneity** is addressed by including multiple datasets with different answer spaces and task formulations, which allows failure patterns to be compared across datasets rather than optimizing for a single benchmark.

- **Transparency of model behavior** is operationalized through reporting not only aggregate measures, but also imbalance-sensitive measures, question-family slices, and concise boundaries in situations where supporting evidence is lacking.

This chapter does not endeavor to supersede existing GI work in the fields of detection, segmentation or video-level robustness. Rather, it supplements that literature with a specialized empirical stratum of **question-answer reliability,** a pre-condition of interactive, clinician-facing MedVQA systems.

3.3 Evaluation Metrics and Statistical Protocol

In this chapter, the model families are compared between the heterogeneous GI MedVQA task regimes. Since these regimes vary in the format of answers (fixed labels vs free text), the structure of labels (nominal vs ordinal) no single measure is sufficient to measure reliability. Chapter 3 thus employs a layered metric bundle, and a conservative reporting protocol: point estimates are only reported when there are matching persisted predictions and evaluation outputs, and comparative claims are constrained by the evidence in the persisted artifacts (Simula n.d.).

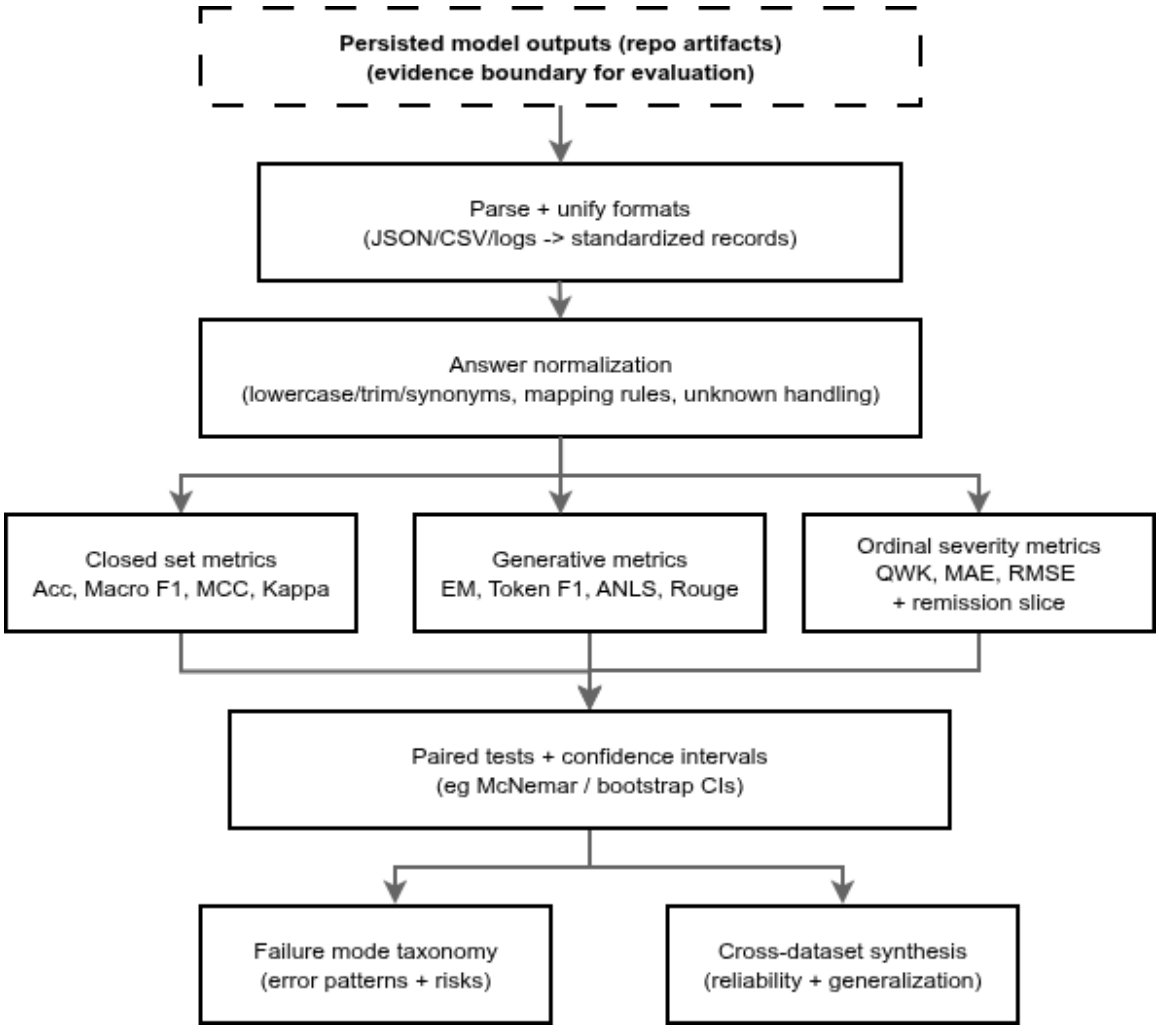

Figure 3.5. Artifact-driven benchmarking workflow

3.3.1 Metric Layers

Chapter 3 employs four metric layers based on task style and clinical risk profile to ensure that there is no single-metric bias.

(A) Closed-set classification and constrained QA

The metrics are applied when outputs are discrete and answer space is discrete (e.g., HyperKvasir class labels, ImageCLEF label IDs, restricted Kvasir-VQA subsets).

- **Accuracy**: overall fraction of correct predictions. Hence, accuracy is used as a measure of correctness rather than of reliability. Precision is interpreted more specifically: it measures the fraction of positive predictions (labels, fields, items, claims) that are correct or relevant. In unbalanced clinical tasks, accuracy and precision have to be interpreted in conjunction with recall, macro-F1, balanced accuracy, ordinal agreement and clinical threshold slices.

- **Macro-F1**: unweighted mean of per-class F1, emphasizing minority-class behavior.

- **Balanced accuracy:** mean of per-class recall (equally weights classes regardless of support).

- **MCC (Matthews correlation coefficient):** correlation-style score robust under imbalance (reported where available in persisted outputs).

- **Cohen's kappa:** chance-corrected agreement measure for categorical outcomes, especially nominal or multiclass label predictions; reported where available in persisted outputs.

- **Imbalance diagnostics (where included in artifacts):** head–tail or rare–common recall gaps computed from classwise recall slices (e.g., thresholds defined in the dataset's local report).

(B) Generative overlap and format fidelity (free-text outputs)

They are applied in cases where a model comes up with natural language responses instead of choosing an answer out of a predefined ontology. They are considered diagnostic fidelity indicators in this chapter, not as adequate evidence of clinical correctness (Smedsrud et al. 2021; Tan and Bansal 2019; Yan et al. 2024).

- **Exact match (EM)**: strict string-level match under the normalization used by the local evaluation scripts.

- **Token-F1:** token-level overlap between generated and reference answers, useful when exact match is too strict.

- **ANLS:** edit-distance-based similarity score used in VQA-style evaluation; implementation details follow the persisted evaluation outputs and scripts used to produce the repository artifacts.

- **BLEU / ROUGE-L / METEOR (where available):** standard NLG overlap metrics reported only when computed in the persisted artifacts.

(C) Ordinal severity and clinically meaningful slices (UC severity as flagship use case)

These metrics are used when labels have an ordinal relationship (e.g., Mayo 0–3) and when the clinical decision boundary matters more than overall accuracy.

- **QWK (quadratic weighted kappa):** chance-corrected agreement measure for ordinal categorical outcomes, it penalizes disagreements according to ordinal distance, so confusing 0↔1 is not equivalent to confusing 0↔3.

- **MAE / RMSE**: absolute and squared error over ordinal labels treated as numeric.

- **Spearman correlation (where available)**: rank correlation for ordinal consistency.

- **Clinical remission slice**: a thresholded binary evaluation derived from ordinal severity labels (in LIMUC artifacts, remission is evaluated as Mayo 0–1 vs 2–3, as indicated in the corresponding table caption).

(D) Uncertainty and comparative significance diagnostics (paired predictions only)

To avoid overinterpreting small point-estimate differences, Chapter 3 uses statistical diagnostics only when paired predictions are persisted for the same evaluation set.

- **Wilson confidence intervals:** used for proportions (e.g., accuracy) when included in the report outputs.

- **Paired McNemar test:** used for paired comparison of two classifiers on the same examples; reported only where persisted counts (n01, n10) and p-values exist in the artifacts.

Scenario type	Primary metrics	Why these metrics

Binary clinical detection	recall/sensitivity, macro-F1, MCC	false negatives and imbalance sensitivity
Multiclass closed-set QA	accuracy + macro-F1 + balanced accuracy	aggregate plus per-class fairness
Free-text QA	token-F1 + ANLS + overlap metrics	lexical similarity with tolerance to paraphrase
Ordinal severity	QWK + MAE/RMSE + remission slice	ordinal penalty and clinical thresholding
Model comparison claims	McNemar + CIs	avoids overinterpreting point estimates

Table 3.5. Metric Selection by Scenario Type

Throughout Chapter 3, N denotes the number of evaluated items recorded in the local report (images or QA rows). Metrics are interpreted primarily within-dataset because answer spaces and task definitions differ across datasets.

3.3.2 Important Evaluation Caveats

The evaluation design of this chapter is purposely conservative. The caveats are as follows to state how the results are to be interpreted and what is not stated.

1. **Cross-dataset scores are not directly comparable.** Different datasets use different answer spaces, label ontologies, and question-family distributions. Thus, Chapter 3 applies primarily cross-dataset comparison to find regularities of failure modes, rather than to put models in an international ranking.

2. **Generative overlap is not clinical correctness.** Lexical similarity and compliance with answer format metrics (token-F1, ANLS, BLEU, ROUGE-L, and METEOR) are used to measure lexical similarity and answer-format compliance. They do not assure generated text to be clinically faithful or visually grounded (Smedsrud et al. 2021; Tan and Bansal 2019; Yan et al. 2024). Generative overlap metrics are addressed in this chapter as diagnostic measures and understood along with failure measures (unknown/OOV behavior, mapping fragility, and question-family brittle).

3. **Raw vs. projected scoring is explicitly separated (diagnostic only).** Other artifacts that have survived are raw scoring of generated outputs and a second projected diagnostic that projects free text into a restricted answer space, then scores. The chapter is based on the artifact semantics:

    - **Raw scoring** is a direct comparison to the canonical targets in the local evaluation convention; in many label-ID tasks, raw free-text output can completely degenerate to near-zero label accuracy due to the failure to present output in canonical label form.

    - **Projected scoring** predictively encodes generated text to the permitted answer set (e.g., canonical labels in a question family) based on the evaluation logic of the repository, and scores the mapped label.

- **Unknown/OOV handling:** if an output cannot be mapped to the allowed answer set, it is treated as unknown/OOV for that diagnostic and contributes to unknown/OOV rates where reported.

Interpretation rule: the scores of format and ontology alignment measured by projected scores in deterministic mapping are not considered as an alternative to primary reliability evidence.

1. **Scenario results are reported only when persisted scenario predictions are available.** Scenario YAML may be provided as protocol context, although the reporting of scenario outcome metrics is only done when scored predictions have been preserved. When configuration is the only available, Chapter 3 reports the configuration and defers scenario outcomes.

2. **Statistical tests are reported only when paired predictions are available.** McNemar tests and confidence intervals need identical predictions on the same samples. The persisted artifacts (these counts and p-values) are reported in Chapter 3 where they are included. The Chapter 3 reports where they are not available include point estimates without significance claims.

3. **Missing metrics are treated as "not computed," not "zero."** In the persisted evaluation results, where a metric is not computed or not calculated (e.g. BLEU/ROUGE not calculated on all runs), the chapter shows this metric as not available and does not make any inferences.

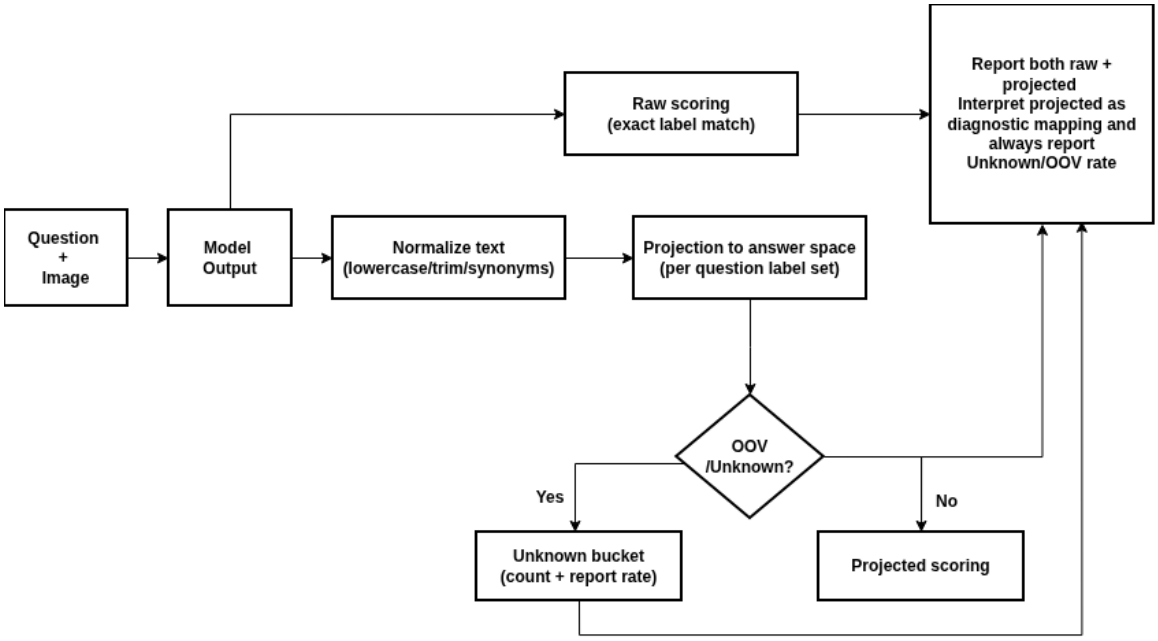

Figure 3.6. Label projection / answer normalization pipeline (raw vs projected + unknown/OOV)

3.3.3 Metric Design Choices Grounded in Prior Work

The metric stack of this chapter adheres to a typical rule in clinical AI evaluation: it should describe the performance in terms of aggregate discrimination and resistance to imbalance and clinically relevant thresholds, as opposed to describing it in terms of one headline score (Simula n.d.).

In the case of GI endoscopy work, this means:

- **macro-F1 and balanced accuracy** are stressed in order to avoid such dominance of majority-class that may conceal clinically relevant failures
- **MCC and kappa** are reported where available because they capture categorical agreement behavior under imbalance, Cohen's kappa is used for nominal or multiclass categorical outcomes, while quadratic weighted kappa is used for ordinal Mayo severity labels.
- **ordinal metrics and clinical slices** (QWK, MAE/RMSE, and remission-style thresholds) are required for UC severity tasks where the distance between grades and the decision boundary are clinically meaningful.

In the case of generative tracks, overlap measures are stored as format and lexical-fidelity diagnostics, though not considered adequate clinical reliability endpoints. In cases where projection/ mapping diagnostics are known, they are explicitly reported as diagnostics to measure ontology alignment and unknown/OOV behaviour.

Lastly, the comparative statistics (e.g., McNemar) are applied as protective measures against overinterpreting small differences, but only in cases where paired evidence is maintained. The chapter thus favors reportable traceable artifact-based reporting as compared to full coverage of metrics.

3.4 Baseline and Existing-Model Results

In Section 3.4, dataset-wise results are reported from the persisted artifacts summarized in Table 3.1 and Appendix A. Since the different datasets specify varying answer spaces (class labels, label IDs per question, or free-text answers) and varying task regimes (classification, closed-label VQA, subset QA), results are viewed in the context of each dataset. To find common patterns of reliance, like sensitivity to imbalance in classes and instability in answer formats, cross-dataset comparisons are made, not to obtain a single overall ranking.

All subsections of the datasets have the same format: (i) the purpose of the dataset in the thesis evaluation design, (ii) the families of models that are compared and why such models are included, (iii) the interpretation of the reported tables, with general metrics, robustness-oriented slices, and paired diagnostics where possible, and (iv) limitations of the results and their implications in the further chapters.

3.4.1 HyperKvasir: 23-Class GI Image Classification

Though not a VQA dataset, HyperKvasir offers a valuable visual grounding stress test of GI imaging (Rieff et al. 2025; Zhang et al. 2023). The dataset here is a 23-class GI image classification that serves as a proxy to the extent of various visual backbones and representations capturing GI categories in realistic data conditions. The main challenge is high long-tail imbalance: in the persisted report, test-set class support lies between 1 and 115 per class (≈115× ratio), i.e. a model can be high-aggregate-accurate even though it fails on uncommon classes that could be clinically significant.

Table 3.6 provides a summary of performance on overall tests of supervised and frozen-feature baselines, and a zero-shot VLM baseline in a label space (as measured in the persisted artifact outputs).

Model	Accuracy	Balanced Acc	Macro-F1	MCC	Kappa
resnet50_supervised	0.8789	0.6266	0.5943	NA	NA
vit_supervised	0.8714	0.5391	0.5242	NA	NA
vit_frozen_logreg	0.8620	0.6130	0.6052	0.8505	0.8504
clip_linear	0.8620	0.5799	0.5721	0.8503	0.8503
blip2_zero_shot_clip	0.0638	0.0529	0.0254	0.0386	0.0303

Table 3.6. HyperKvasir Overall Test Metrics

Table 3.6 summarizes overall test performance for supervised and frozen-feature baselines, alongside a zero-shot VLM baseline projected into a label space (as recorded in the persisted artifact outputs).

The supervised and frozen-feature pipelines establish a solid reliability foundation on aggregate measures whereas the zero-shot VLM baseline fails in this closed-label classification environment. Nevertheless, combined measures do not work in a long-

tail regime. Specifically, the focus will be on balanced accuracy and macro-F1 since they consider the minority classes more than the raw accuracy.

Long-tail robustness can be explicitly quantified by reporting a slice of rare-class and common-class recalls (as in the persisted report), as well as a head-tail gap in recall.

Model	Rare-class recall (support <= 5)	Common-class recall (support >= 90)	Common-minus-rare gap
resnet50_supervised	0.1595	0.9375	0.7779
vit_supervised	0.0000	0.9432	0.9432
vit_frozen_logreg	0.1714	0.9120	0.7406
clip_linear	0.0476	0.9091	0.8615
blip2_zero_shot_clip	0.0000	0.0471	0.0471

Table 3.7. HyperKvasir Imbalance Robustness Slices

Every high-accuracy model has a high level of head-tail asymmetry: common-class recall is high and rare-class recall is low. It is interesting to note that some of these models perform very well in recalling common classes with near zero recall on rare classes. This explains the need to have long-tail assessment slices to be risk-conscious. It also encourages the subsequent chapters to consider the minority-class behavior as a fundamental reliability condition, not that high overall accuracy means safe performance.

Table 3.8 indicates the selected McNemar tests to test the hypothesis that the observed differences between model variants are only due to fluctuations of the order of noise on the same test examples where paired predictions are present in the persisted artifacts.

Pair	n01 (A wrong, B right)	n10 (A right, B wrong)	p-value
vit_frozen_logreg vs clip_linear	68	68	0.931666
vit_frozen_logreg vs blip2_zero_shot_clip	21	871	< 1e-6
clip_linear vs blip2_zero_shot_clip	19	869	< 1e-6

Table 3.8. HyperKvasir Pairwise McNemar Tests

In the continued paired comparisons, vit_frozen_logreg and clip_linear do not indicate any significant difference as the counts of disagreements are symmetric with a

large p-value. In comparison, both are distinctly distinctly different to the zero-shot baseline, with the number of disagreements heavily asymmetric, and the p-values very small. These tests confirm the qualitative conclusion that is already present in Table 3.6: in the framework of this dataset and the existing evidence, the use of supervised or frozen discriminative pipelines is better than the zero-shot one.

Limitations and implications. Here, HyperKvasir is a GI visual-grounding stress test and not a direct VQA benchmark. Its main worth is thus diagnostic. In particular, it demonstrates that good aggregate performance may be accompanied by low recall on infrequent classes, and that sound GI perception in the long-tail imbalance is a bottleneck which needs to be checked explicitly. This encourages subsequent chapters to maintain good visual grounding elements and to consider class-imbalance mitigation as a requirement of clinically aligned behavior.

3.4.2 ImageCLEF MEDVQA-GI 2023: Closed-Label GI VQA

ImageCLEF MEDVQA-GI 2023 offers a closed-label VQA environment, where every question is linked with a limited answer ID space (Simula Datasets n.d.; Hu et al. 2021). This allows deterministic testing and is ideal to answer the main reliability question of the thesis: how do fine-tuned multimodal VQA models perform in comparison to raw zero-shot generative VLM results when constrained with labels? In this chapter, it is assessed the persisted validation split (N = 7,332) that was documented on the local report artifacts.

Table 3.9 provides general validation figures of a fine-tuned transformer VQA model and two zero-shot forms: “raw” (direct outputs) and “projected” (diagnostic mapping of outputs to the allowed answer space, as defined in Section 3.3).

Model variant	N	Accuracy	Balanced Acc	Macro-F1	MCC	Kappa
vilt_finetune	7,332	0.9089	0.5853	0.5823	0.8876	0.8875
qwen2_5_vl_zeroshot_raw	7,332	0.0007	0.0433	0.0007	-0.0696	-0.0626
qwen2_5_vl_zeroshot_projected	7,332	0.0670	0.0899	0.0379	-0.0296	-0.0278

Table 3.9. ImageCLEF MEDVQA-GI 2023 Validation Metrics

The fined tuned VQA model has a good performance in the closed-label accuracy, whereas the raw zero-shot generated outputs have a near score of zero in the label-ID evaluation format. Projection increases the zero-shot diagnostic score, which means that a portion of the generated outputs can be projected into the label space, but the projected accuracy and macro-F1 is still way lower than the fine-tuned baseline. The zero-shot variants have negative MCC/kappa values, which are in agreement with the systematic mismatch in the constrained label-ID evaluation format, and are not random.

Table 3.10 presents family-level aggregates to not conceal question-family brittleness. This is significant since there are families (e.g., procedure or attribute) which might demand stronger adherence to ontology than others.

Question family	Rows	ViLT acc	Qwen raw acc	Qwen projected acc	ViLT macro-F1	Qwen projected macro-F1
attribute	1,600	0.8488	0.0000	0.0225	0.5153	0.0340
binary/boolean	2,800	0.9339	0.0004	0.1168	0.9222	0.0906
count	1,200	0.9008	0.0017	0.0400	0.3950	0.0158
location	1,332	0.9092	0.0015	0.0601	0.3115	0.0205
procedure	400	0.9975	0.0000	0.0000	0.9969	0.0000

Table 3.10. Family-Level Aggregates on ImageCLEF Validation

There are two patterns evident. First, the fine-tuned model is always robust across families in terms of accuracy, whereas macro-F1 differs significantly across families (particularly count/location) which suggest that even good aggregate accuracy can mask intra-family imbalance or challenge. Second, the raw outputs of the zero-shot model are near zero across families and projection achieves partial recovery in the persisted artifacts on binary/boolean questions only, with little recovery on attribute, count and location and no recovery on the procedure family. This implies that projection is able to salvage some of the discrepancy in output format, but they do not remove the underlying reliability gap within structured question families. Table 3.11 displays the questions that in the persisted validation results showed the largest raw-to-projected improvement in the accuracy measures.

Question	Raw acc	Projected acc	Absolute gain
Is there a green/black box artefact?	0.0000	0.5475	+0.5475
Are there any instruments in the image?	0.0000	0.1800	+0.1800

Where in the image is the abnormality?	0.0000	0.1400	+0.1400
What color is the abnormality?	0.0000	0.0800	+0.0800
How many polyps are in the image?	0.0025	0.0600	+0.0575

Table 3.11. Largest Qwen Lexical-Projection Gains (Validation Accuracy)

Those examples suggest that deterministic mapping can provide a significant contribution to the accuracy of measurable labels of some prompts, which implies that some raw outputs carry some recoverable information even when presented in a non-predicted label scheme. Nonetheless, as mentioned in Section 3.3, projection is also a diagnostic tool, and can be overly correct when the apparent match is superficial or founded on lexical similarity only. The most significant reliability finding of this chapter is thus the consistent discrepancy of Table 3.9 and Table 3.10: the zero-shot results are still significantly lower than the fine-tuned closed-label baseline on major question groups.

Limitations and implications. In this subsection, the comparisons between a fine-tuned transformer baseline and one of the zero-shot models in the persisted evaluation setup are made. It does not purport to generalize to all VLMs or all GI datasets. Further, due to the lack of paired statistical tests in this data in the provided tables, this chapter does not use the word significant here and treats differences as the large observed gaps amid the available evidence. Its practical implication is that answer-space governance (constrained decoding, ontology compliance) and probable supervision or adaptation is needed to achieve reliable performance in label-ID GI VQA settings, which lead to the controlled generative design choices presented in Chapter 4.

3.4.3 Kvasir-VQA: Subset Reliability and Answer-Format Stability

Kvasir-VQA is a colonoscopy-based VQA tool and the core of the supervisor-guided empirical pathway to this dissertation (Safwan et al. 2025; Gautam et al. 2024). The artifacts that were persisted to this chapter are (i) structured subset assessments of yes/no questions, attribute-style questions, (ii) a reformatted phase III generative snapshot designed to store runtime status. Two of the reliability themes that are important to GI MedVQA are addressed by this subsection: (1) the strength of constrained/fusion pipelines on structured question families, and (2) the weakness of unconstrained generation when the output format is not controlled.

Model	N	Accuracy	Balanced Acc	Macro-F1	MCC	Unknown rate

resnet_gru_m1_yesno	443	0.986456	0.973673	0.964953	0.930163	0.0000
vit_bertlite_m2_yesno	443	0.950339	0.906593	0.878126	0.759432	0.0000
blip2_zeroshot_yesno	443	0.893905	0.509376	0.492332	0.086138	0.0000
blip_vqa_base_yesno_forced_choice	12,267	0.518301	0.514587	0.502888	0.030894	0.0000
blip_vqa_base_yesno_freegen	500	0.000000	NA	0.000000	NA	1.0000

Table 3.12. Kvasir-VQA Yes/No Results

There are two different patterns. First, limited fusion models (resnet_gru_m1_yesno, vit_bertlite_m2_yesno) are highly accurate and balanced on this subset that has been persisted, which means that structured yes/no questions can be processed reliably when the output space is controlled. Second, the zero-shot model (blip2_zeroshot_yesno) has high accuracy, yet it has almost-chance balanced accuracy and low MCC, which is predictable given the issue of majority-class bias in an imbalanced yes/no distribution- an instance of why macro-f1, balanced accuracy and MCC are needed as complements to raw accuracy. Lastly, the free-generation (blip_vqa_base_yesno_freegen) variant fails under this assessment format with unknown-rate = 1.0, which demonstrates that there is full nonadherence to the persisted scoring rules in the answer-format.

Table 3.13 reports results for an attribute subset, which is more challenging than yes/no because it typically expands the answer vocabulary and increases ambiguity.

Model	**N**	**Accuracy**	**Balanced Acc**	**Macro-F1**
resnet_gru_m1_attribute	352	0.670455	0.376936	0.367341
vit_bertlite_m2_attribute	352	0.656250	0.374696	0.355586

Table 3.13. Kvasir-VQA Attribute Subset

Accuracy is moderate, but the balance and macro-F1 are significantly lower, meaning that the attribute subset is more imbalanced and/or confuses the classes than the yes/no subset. This is in line with the bigger picture of reliability perspective taken in this dissertation: although constrained pipelines may work well on the yes/no family of questions that are dominating, more challenging question families can still be bottlenecks and thus have to be analyzed separately and not be extrapolated to the easiest ones.

In this sub section we have subset level evidence as opposed to a complete Kvasir-VQA benchmark sweep. The yes/no (N = 443) and the bigger forced-choice assessment (N = 12,267) cannot be directly compared as a controlled experiment, as they vary in the size of the evaluation set, and may also vary in the way the subsets were made. However, the reliability message is the same as the persisted artifacts: (i) constrained/fusion pipelines can be extremely powerful on structured subsets, and (ii) unconstrained generation may fail under closed-label evaluation because of the instability of answer formats. These findings drive the design posture adopted in later chapters: maintain constrained pathways to high-reliability sub-questions and only add generative answering in the face of explicit output control (constraints, mapping, or abstention behavior).

3.4.4 Kvasir-VQA-x1: Large-Scale Generative Reasoning Benchmark

The largest QA environment is Kvasir-VQA-x1, which is the strongest stress test in this dissertation for open-ended generative answering under greater question variety and difficulty (Gautam et al. 2025a; Simula n.d.; Appendix A, Artifact A5). In contrast to closed-label tasks (where responses are chosen among a fixed ontology), Kvasir-VQA-x1 tests free-text generation. Consequently, reliability in such an environment is not solely based on the quality of the underlying visual reasoning, but also on the stability of the answer format, lexical normalization and ontology alignment. All of them can degenerate even when a model generates a text that seems real.

In this subsection, three integrative perspectives of model behavior are shown:

1. **Generative overlap diagnostics** quantify lexical similarity between the generated output and reference responses according to the evaluation conventions in the persisted artifacts.

2. **Closed-set style diagnostics and baselines** demonstrate how mapping and answer-space governance can affect apparent performance, such as failure modes like high OOV (out-of-vocabulary) or unknown mapping rates.

3. **The complexity-level breakdown** is a descriptive diagnostic of the changes in token-level overlap with question complexity as observed in the persisted summaries.

Generative adaptation and overlapping effects.

Model	EM	Token-F1	ANLS	BLEU	ROUGE-L	METEOR	Count
medgemma_lora_original	0.000000	0.508473	0.340755	NA	NA	NA	15,955
medgemma_zeroshot	0.000063	0.213080	0.017498	0.033341	0.158501	0.141180	15,955
llava_zeroshot	0.000000	0.212437	0.007032	0.025760	0.163942	0.150097	15,955
qwen2_5_vl_zeroshot	0.000000	0.172788	0.000000	0.017084	0.123496	0.187288	15,955

Table 3.14. Kvasir-VQA-x1 Generative Metrics

Exact match (EM) is close to zero across all runs. This is consistent with the use of strict Precise match (EM) is near zero in all the runs. This is in line with strict string matching in free-text generation whereby a slight change in lexicon can block a perfect match even though there might be a lot of similarity in meaning. Token-F1 and ANLS offer less intolerant diagnostics and exhibit a pronounced response to adaptation: the LoRA-adapted run significantly boosts token-level overlap with respect to the zero-shot runs in the persisted artifacts. BLEU/ROUGE/METEOR NA is where the metric has not been calculated in the persisted outputs and the metric is not inferred.

Reliability boundary. Such overlap measures capture the fidelity of the surface and the adherence of the answer format, but do not ensure clinical accuracy or support of the visual grounding (Section 3.3)(Smedsrud et al. 2021; Tan and Bansal 2019; Yan et al. 2024). They are, therefore, considered diagnostic cues and analyzed along with mapping/OOV signals and class-balanced behavior.

Closed-set style baselines and mapping fragility

To reveal the influence of answer-space governance over performance in appearance the persisted artifacts are the new diagnostics and baselines that project the outputs into restricted spaces or assess other non-generative methods.

Model	N	Accuracy	Balanced Acc	Macro-F1	Notes
fusion_tfidf_vit_logreg	5,893	0.814865	NA	0.150749	fusion baseline

text_yesno_tfidf_logreg	1,540	0.777922	NA	0.777291	yes/no-specific
vlm_zeroshot_label_mapped	15,955	0.561642	NA	0.005817	OOV rate 0.973
text_topk_tfidf_logreg	4,252	0.422389	0.235698	0.204103	top-3 0.7408
image_resnet50_logreg	5,952	0.233535	NA	0.008556	image-only
text_bert_classifier	9,148	0.158942	0.006654	0.002327	weak generalization
image_vit_logreg	4,252	0.020461	NA	0.006352	image-only

Table 3.15. Kvasir-VQA-x1 Mapped/Baseline Diagnostics

In this case, two reliability signals are significant:

1. **Moderate mapped accuracy can be misleading.** There is a high accuracy of the mapped VLM diagnostic with a value of above 0.5, macro-F1 is close to zero, and the OOV rate is very high. This indicates a gross incompatibility in the answer space: not many of the outputs can be relocalized to the desired label set, and not the relocalized subset itself shows balanced behavior. In accordance with the statements in Section 3.3, it is thus the responsibility of mapped accuracy to be assessed as a format and ontology diagnostic but not as a sign of clinically reliable reasoning.

2. **Subset-specific baselines can look strong due to question-family skew.** The yes/no-specific baseline indicates high macro-F1 in the limited subset, which supports the necessity to report the results in terms of task regime and to not use the easy subsets to generalize to the reliability.

Complexity effects (descriptive diagnostic)

Complexity	llava_zeroshot	medgemma_zeroshot	qwen2_5_vl_zeroshot
1	0.151298	0.145365	0.079875
2	0.217874	0.216663	0.171684
3	0.271474	0.280927	0.271952

Table 3.16. Token-F1 by Complexity Level

The level of complexity of the persisted summaries increases with token-F1. This trend is only reported descriptively, as increased complexity can have longer outputs, or repetitive lexical cues that expand overlap scores but do not necessarily indicate more

grounded reasoning. That is why the trends of complexity-level overlap are to be viewed alongside the OOV behavior and family-level diagnostics.

Limitations and implications. Kvasir-VQA-x1 indicates a generic generative MedVQA trend in the repository evidence: strict EM collapses, token-level overlap enhances with adaptation, and deterministic mapping can achieve moderate accuracy despite an OOV behavior and biases in the answer space. The pragmatic conclusion to Chapter 4 is that the introduction of generative capability along with answer-space governance, such as controlled decoding, normalization or mapping protection, and abstinence behaviour should be considered, as opposed to generative capability serving as an unconstrained path to answers.

3.4.5 LIMUC: UC Severity Reliability (Flagship Clinical Axis)

In Chapter 3, LIMUC offers the most clinically direct evidence because it deals with a UC severity grading task (Mayo 0–3), an ordinal task with clear clinical stakes (Polat et al. 2022; Stidham et al. 2019; Ozawa et al. 2020; Yao et al. 2023; Takenaka et al. 2023; Appendix A, Artifact A3). In this context, accuracy cannot encompass reliability. Ordinal distance is important, as confusion between 1 and 0 is not the same as confusion between 0 and 3, and the behavioral threshold of clinical importance, e.g., remission versus active disease, has to be studied more explicitly. To this end, ordinal agreement (QWK), the size of errors (MAE/RMSE) and a remission-oriented slice are highlighted in this subsection besides the conventional measures of classification.

Ordinal and classification metrics

Model	Accuracy	Balanced Acc	Macro-F1	QWK	MAE	RMSE
finetune_resnet50	0.753855	0.695008	0.682889	0.835097	0.256821	0.528545
finetune_vit_or_swin	0.727165	0.673848	0.672142	0.806259	0.287070	0.564888
vit_frozen_logreg	0.689798	0.641650	0.618454	0.758806	0.348161	0.654848
clip_linear_baseline	0.679122	0.635709	0.602016	0.745502	0.367734	0.687112
resnet50_frozen_logreg	0.619217	0.542258	0.533958	0.679280	0.434757	0.742299
vlm_zero_shot_mayo	0.548636	0.250000	0.177135	0.000000	0.698695	1.155727

Table 3.17. LIMUC Overall Test Metrics

Finely-tuned visual backbones are more accurate and have better macro-F1 and ordinal agreement (QWK) and smaller error magnitude (MAE/RMSE), so they are better at grading severity more consistently and with fewer large ordinal errors. In comparison,

the zero-shot severity-prompt baseline exhibits significantly lower balanced scores and fails on ordinal agreement (QWK = 0.0) when using the persisted evaluation condition, suggesting that ordinal consistency is not achieved in this zero-shot prompting conditioning.

Clinical threshold slice (remission vs active disease)

Since clinical workflows tend to be based on threshold-based reasoning, such as between remission and active disease, the artifacts that persist also comprise a remission-oriented slice, as defined in the LIMUC report.

Model	Remission accuracy	Sensitivity	Specificity	Remission F1
finetune_resnet50	0.947805	0.967603	0.855219	0.968300
finetune_vit_or_swin	0.937722	0.968323	0.794613	0.962433
vit_frozen_logreg	0.902135	0.917207	0.831650	0.939182
resnet50_frozen_logreg	0.886714	0.917927	0.740741	0.930317
clip_linear_baseline	0.886714	0.895608	0.845118	0.928705
vlm_zero_shot_mayo	0.823843	1.000000	0.000000	0.903415

Table 3.18. LIMUC Remission Slice Metrics

Sensitivity and specificity of this slice are kept at manageable levels by tuned and frozen-feature baselines. In comparison, the zero-shot baseline exhibits a critical threshold failure: the sensitivity is 1.0 and the specificity is 0.0, that is, there is no ability to differentiate the negative category when using this definition of operational. It is precisely this type of high-risk behavior that accuracy can mask, and an example of why clinical slice analysis is required in severity-oriented MedVQA.

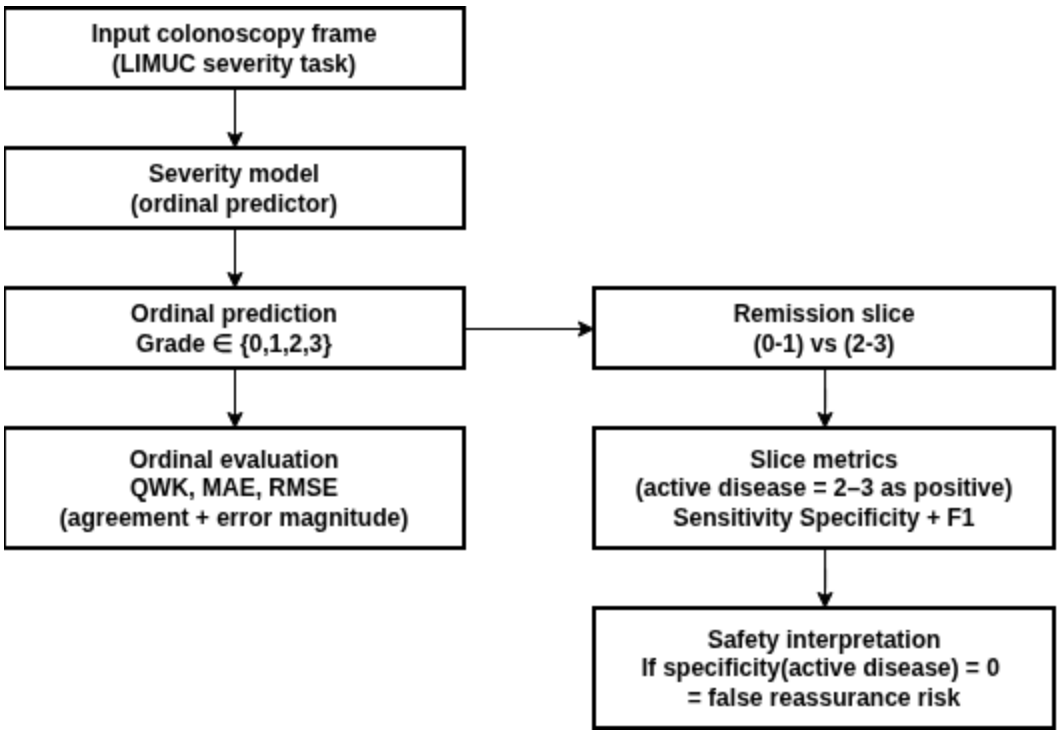

Figure 3.7. UC severity evaluation workflow (LIMUC)

Per-class bottlenecks

To make minority class behavior explicit, the persisted artifacts also include a best-per-class summary.

Mayo class	Support	Best model	Best F1	Best recall
0	925	finetune_resnet50	0.852516	0.796757
1	464	finetune_resnet50	0.683859	0.771552
2	177	finetune_resnet50	0.552326	0.536723
3	120	finetune_vit_or_swin	0.683544	0.675000

Table 3.19. LIMUC Per-Class Best Model Summary

Even with the best-performing models, support is very disproportionate between the Mayo classes, with the worst performance in the intermediate-severe classes, especially in Mayo 2. This justifies the thesis motivation that minority severe classes continue to be the primary bottleneck to reliability and thus explicitly need to be addressed in both evaluation and method design.

Limitations and implications. In Chapter 3, LIMUC reports the most clinically actionable evidence: under the persisted evaluation configuration, the supervised domain-tuned pipelines offer significant improvements in ordinal and clinical-slice scores over the zero-shot baseline. The Chapter 4 practical implication is straightforward: UC severity generation needs to conserve ordinal structure, have clinically meaningful threshold behavior, particularly specificity, and cannot rely on raw zero-shot generation without protection.

3.5 Cross-Dataset Synthesis and Findings

This part summarizes the results at the dataset level into thesis level results that connect to the research questions. Due to the difference between the datasets in terms of answer space and task regime, cross-dataset comparisons serve more to reveal some common reliability hierarchies and failure modes, than to assert a global ranking.

3.5.1 Comparative Reliability: Constrained vs Zero-Shot

Dataset	Strongest constrained/tuned result	Zero-shot/open baseline result	Absolute gap
HyperKvasir	resnet50_supervised acc 0.8789	blip2_zero_shot_clip acc 0.0638	-0.8151
ImageCLEF MEDVQA-GI 2023	vilt_finetune acc 0.9089	qwen_projected acc 0.0670	-0.8419
Kvasir-VQA yes/no subset	resnet_gru_m1 acc 0.9865	blip2_zeroshot_yesno acc 0.8939	-0.0926
LIMUC severity	finetune_resnet50 acc 0.7539	vlm_zero_shot_mayo acc 0.5486	-0.2052
Kvasir-VQA-x1 generative	medgemma_lora token-F1 0.5085	qwen_zeroshot token-F1 0.1728	+0.3357 (adaptation gain)

Table 3.20. Reliability Gap Snapshot Across Datasets

The evidence in the repository repeatedly suggests that the reliability of the structured GI tasks is still constrained/supervised pipelines, whereas the raw zero-shot generation is less robust, and less format-stable. The Kvasir-VQA-x1 row is a summary of a generative overlap diagnostic (token-F1) and not accuracy; it demonstrates that adaptation enhances overlap fidelity but fails to overcome answer-space governance and grounding problems found in the dataset-wise analysis.

Result for RQ2. Within the sustained evidence in this repository, constrained/supervised methodologies are the most secure fundamental route to GI MedVQA and zero-shot transfer is always inferior to stringent testing criteria.

3.5.2 Dominant Failure Modes (RQ3)

Failure mode	Evidence in this chapter	Practical implication
Head-tail imbalance collapse	HyperKvasir rare recall near zero for multiple models	aggregate accuracy can mask clinically important misses
Lexical drift / non-answer generation	Kvasir-VQA freegen unknown-rate 1.0	requires constrained decoding and output guards
OOV mapping fragility	Kvasir-VQA-x1 mapped VLM acc 0.5616 but macro-F1 0.0058 with OOV 0.973	mapped accuracy alone can be misleading
Question-family brittleness	ImageCLEF: procedure/attribute families collapse for zero-shot raw/projected	per-family reporting is mandatory
Clinical threshold blind spots	LIMUC zero-shot remission specificity 0.0	unsafe threshold behavior without supervision
Scenario evidence gap	reformatted tree provides scenario config without scored outputs	scenario-level failure conclusions are deferred until outputs are persisted

Table 3.21. Observed Failure Taxonomy

These failure modes justify why single-score reporting cannot be used to clinical MedVQA(Simula n.d.). There are data-driven (long-tail imbalance) and format-driven (unknown/OOV and mapping fragility) and clinically risk-driven (threshold failures and ordinal inconsistency) failures. This taxonomy is a motivating factor to the design of risk-control and evidence-governance in subsequent chapters.

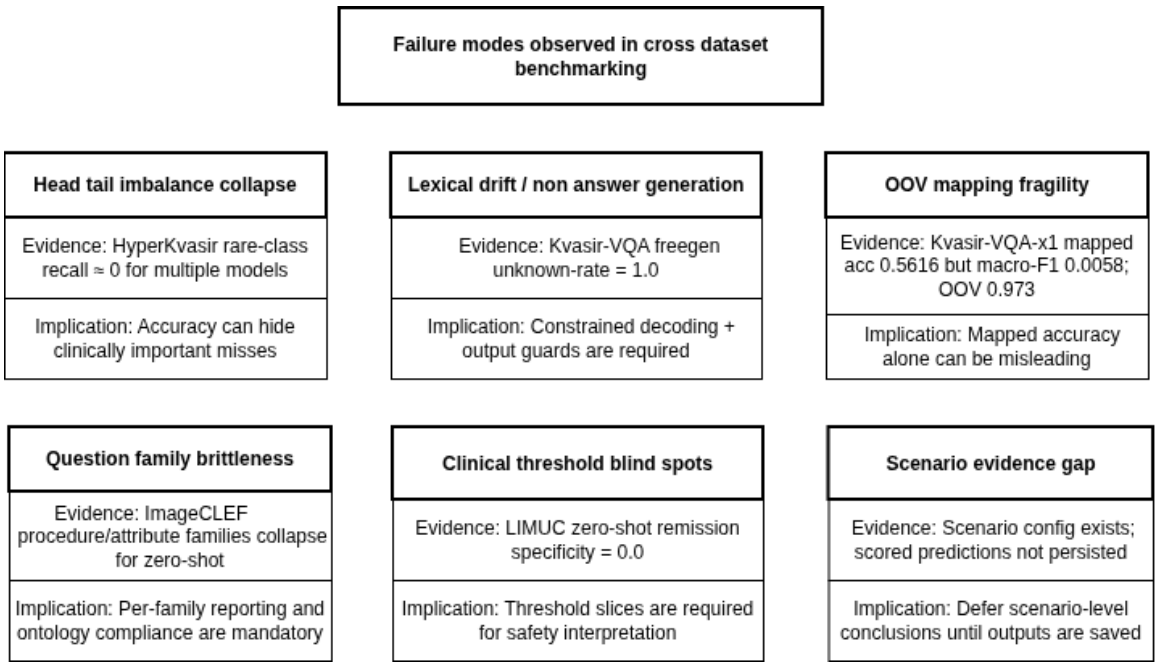

Figure 3.8. Failure taxonomy tree

3.5.3 Severity Robustness (RQ4)

The LIMUC results present the best empirical evidence of RQ4. The major finding is that the severity-focused MedVQA should be assessed in terms of ordinal and threshold-congruent measurement, and not in terms of accuracy. When tuned models are used in the persisted evaluation configuration, they exhibit improved ordinal agreement (QWK) and more stable remission-threshold behavior than the zero-shot baseline, whereas per-class results indicate that underrepresented moderate-to-severe classes continue to be the main bottleneck.

3.5.4 Statistical Stability

In the presence of paired predictions and McNemar diagnostics, where paired predictions and McNemar diagnostics are persisted and reported, such as in particular blocks of data that contain n01/n10 counts and p-values, the data indicate that some of the larger observed differences cannot be due to noise-level fluctuations only. Where paired tests are absent in the chapter tables in dataset blocks, the difference values are reported in a more conservative way, as descriptive observed differences as opposed to formal significance claims.

3.5.5 Threats to Validity and Boundaries

Threat	Potential bias	Mitigation applied
Cross-dataset task heterogeneity	direct metric comparison may be invalid	comparisons are primarily within dataset/task
Incomplete artifact parity	some runs have training logs but missing validation preds	explicitly marked as unavailable; no fabricated comparisons
Label projection inflation risk	projected text may match labels lexically without semantic correctness	projected scores reported as diagnostic, not final clinical score
Scenario artifact incompleteness	no persisted reformatted scenario predictions/metrics	scenario claims removed from quantitative synthesis
Legacy vs reformatted pipeline differences	potential metric provenance mismatch	chapter constrained to Prototyping_reformat sources only

Table 3.22. Threats to Validity and Mitigation

The main methodological threat in this chapter is the excessive interpretation of unfinished artifacts. To cope with this, the chapter uses a provenance-first approach: all assertions are directly related to maintained tables, and diagnostics, e.g. projection or mapping, are explicitly defined as diagnostics and not as an assertion of clinical soundness.

3.5.6 Positioning Against Broader MedVQA Findings

These empirical trends align with the general results of MedVQA discussed in Chapter 2: constrained or fine-tuned pipelines remain more predictable than naive zero-shot generation on structured clinical problems, and advances in overlap-based generative measures do not necessarily suggest that they are guided by grounded or clinically faithful reasoning (Smedsrud et al. 2021; Tan and Bansal 2019; Yan et al. 2024) . The Chapter 3 contribution is not thus a state-of-the-art claim, but a reproducible reliability map of GI MedVQA in a single envelope of artifact. Specifically, it recognizes format instability, mapping fragility, question-family brittleness, and threshold failures that need to be mitigated prior to deployment to clinicians.

3.6 Position After Chapter 3

Chapter 3 builds an empirical ground of the dissertation that is reproducible:

1. GI MedVQA performance is very sensitive to the type of task and the answer format, closed-label and free-text conditions are different and need different assessment strategies.

2. According to the continued evidence within this repository, constrained and supervised approaches are the most certain foundation to closed and ordinal clinical activities.

3. Naive zero-shot generation is not adequate as a standalone clinical answer pathway on the evaluation conventions and ontology requirements.

4. Severity-focused evaluation must remain central, particularly ordinal agreement and clinically meaningful threshold slices such as remission specificity.

5. The generative capability is useful only in the sense of being amenable to explicit governance, such as constraints and normalization, unknown or OOV processing, abstention behavior, and, later in the book, evidence-based grounding of higher-level queries.

The results of these results then lead to Chapter 4 that constructs a controlled generative pipeline of UC severity that maintains ordinal consistency and adds safeguards to output stability and then moves to evidence-conscious responding.

Chapter 4.

Generative Vision-Language Modeling for Ulcerative Colitis Severity Assessment

4.1 Chapter Overview and Methodological Rationale

Chapter 3 has laid down the main empirical conflict which drives the current chapter. In various GI-endoscopy environments, constrained supervised models proved to be more effective compared to naive zero-shot vision-language prompting. This trend was particularly significant in ulcerative colitis (UC) severity grading on LIMUC, in which the activity is clinically relevant, ordinal in nature, and the imbalance of classes is sensitive. The methodological question then becomes not whether or not it is possible to generate, but how a generative model can be modified to retain the discipline of a classifier and the flexibility of a language-based interface.

The answer to this question can be found in Chapter 4 which creates a severity-based, controlled pipeline to Mayo 03 scoring of colonoscopy frames. It does not seek free narrative generation. Rather, generative modeling is viewed as a constrained decision process: the model must generate a highly constrained score output, the model predictions are interpreted based on explicit rules, and the model behaviour is assessed based on a strong supervised baseline on the same data split and reporting regimen. The significance of this framing is that this dissertation is interested in clinically applicable multimodal systems, not in open-ended text production per se.

The chapter has four objectives:

1. To define a reproducible UC severity task on LIMUC using fixed Mayo 0–3 labels and a fixed split structure.

2. To establish a strong supervised reliability anchor against which generative methods can be judged fairly.

3. To implement parameter-efficient adaptation of a vision-language model using LoRA so that score generation is learned rather than merely prompted.

4. To assess the resulting system with ordinal, class-balanced, and clinically interpretable metrics as well as to make the failure modes explicit.

The rest of the chapter is based on this logic. Section 4.2 delineates the dataset and the scope of claims made. Section 4.3 introduces the proposed pipeline and describes the purpose of each of the modeling layers. The evaluation protocol is formalized in section 4.4. The internal and external results are reported in Section 4.5. Section 4.6 explains those findings within the context of the overall argumentation of the dissertation, whereas Section 4.7 identifies the limitations which limit the claims of the chapter. Section 4.8 wraps up by placing the resultant severity module as the upstream element to the evidence-based wrapper, introduced in Chapter 5.

4.2 Dataset, Clinical Task, and Scope

Our solution is not universal. Rather than trying to address general GI MedVQA in a single step, the chapter explores a task which is clinically focused and measurable, the classification of the Mayo endoscopic severity score from a single image of colonoscopy. This makes the experiment more affordable in terms of the question being asked and allows a comparison of discriminative versus generative approaches with the same evidence.

4.2.1 LIMUC as the Primary Evidence Base

This chapter is mostly based on LIMUC, a database of images of the UC with a Mayo score of 0 to 3. They are ordinal severity scores, rather than a nomenclature. This is why mistakes are not as critical: it is not the same to confuse Mayo 0 with Mayo 1 like it would be to confuse Mayo 0 with Mayo 3. It is one of the reasons why Chapter 4 focuses on ordinal measures, such as quadratic weighted kappa (QWK) and standard measures, such as accuracy and F1.

LIMUC preparation is performed using the notebook listed in Appendix A, Artifact A7, which generates the metadata tables and manifests used in this chapter. The snapshot used in this study, metadata_enriched.csv has 11,276 frames.

Split	Frames
Train	8,669
Validation	921
Test	1,686

Table 4.1. LIMUC Split Distribution

Mayo class	Frames

0	6,105
1	3,052
2	1,254
3	865

Table 4.2. LIMUC Mayo Class Distribution Across All Splits

The characteristics of this distribution affect the design. First, the majority of the dataset are classes 0 and 1, while classes 2 and 3 are under-represented. Second, the task is ordinal, so it's likely that adjacent classes are confused rather than randomly misclassified. Third, the data is large enough to train a supervised model, but it is also unbalanced, so that cannot be interpreted as overall accuracy. Consequently, the proposed method is assessed using balanced and ordinal metrics and the generative lane is trained with balanced sampling rather than simple frequency-based training.

4.2.2 Scope Boundary for Chapter 4 Claims

The major claims in Chapter 4 are deliberately limited to internal LIMUC evaluation using the fixed split and reporting protocol as outlined in this chapter. The focus of the comparison is therefore not on arbitrary best runs from different configurations; it is an apples vs apples comparison between a supervised baseline family and a generative adaptation family on the same internal task.

This is important because several other tasks in the repository, such as Kvasir-VQA, Kvasir-VQA-x1 and ImageCLEF MEDVQA-GI, are relevant to the dissertation, but not necessary to establish the main claim of Chapter 4. They are mostly used as a comparison and a context, especially in Chapters 2 and 3. However, the Chapter 4 remains focused on a single clinically relevant task (the severity task) so that the methodological innovation can be evaluated.

The reporting policy is the same. The main objective for optimization and reporting is QWK of internal LIMUC mode1/test. The remaining metrics (accuracy, macro-F1, balanced accuracy, mean absolute error (MAE), root mean squared error (RMSE), and parse rate) are reported but they don't take precedence over the ordinal agreement task. This makes sense because the Mayo task is ordinal and the performance on this task should not solely be assessed by accuracy.

4.2.3 External HyperKvasir UC Proxy Stress Test

This chapter includes an external-only stress test, using a HyperKvasir-derived UC proxy set. The purpose of this test is not to make an absolute claim of generalization. It is to assess how the benefits seen on LIMUC data hold up to a new data source, label compatibility, and output set.

The external protocol uses metadata_hyperkvasir_uc_proxy_mayo_floor.csv, which is a floor mapping of interval labels. Specifically, interval findings are mapped as 0-1 -> 0, 1-2 -> 1, and 2-3 -> 2. This approach can be used for robustness testing but it is not the same as Mayo annotation. So, the external set is a stress test rather than a final, clinical gold standard.

Mayo proxy class	Frames
0	35
1	212
2	471
3	133

Table 4.3. External HyperKvasir UC Proxy Distribution

The external set comprises 851 frames. Two points need to be made. First, the class distribution is very different from LIMUC, particularly for low-severity classes. Second, the mapped labels in the external dataset introduce uncertainty in the label space. As such, the external results are viewed as an indication of limitation and domain shift, but not as the foundation for model selection.

4.3 Proposed Severity-Oriented Pipeline

The Chapter 4 pipeline can be roughly described as:

dataset curation and split freezing -> supervised and generative baseline construction -> controlled severity prediction -> statistical evaluation -> error analysis

The guiding principle is that the pipeline should bring the gap between open multimodal generation and rigorous clinical scoring closer to zero. The pipeline is thus layered with each layer performing a specific function in disentangling improvement from the effects of the prompt.

4.3.1 Data Preparation and Split Freezing

The initial layer is the data preparation layer, which is implemented in the LIMUC preparation notebook listed in Appendix A, Artifact A7. This is not just about collecting images and labels, but also about specifying a task. The notebook generates metadata tables, split assignments, label maps and a split hash so that any subsequent experiment can be traced to a particular state of the data.

This is a crucial step in the methodology because downstream comparisons would be hard to justify if the composition of the training, validation and test sets were to change. In a dissertation, reproducibility isn't merely a feature, but part of the evidence of the chapter. The chapter ensures that the metadata snapshot and split assignment are frozen before the models are compared, thus preventing a common pitfall of multimodal experiments, where model differences are confounded with inadvertent preprocessing changes.

4.3.2 Supervised Reliability Anchor

The second layer provides a strong supervised anchor. This includes frozen-encoder baselines and fine-tuned baselines. The frozen-encoder baselines are implemented in:

- Prototyping_reformat/DatasetAnalysis/LIMUC/1_frozen_encoders/resnet50_frozen_logreg.ipynb (Appendix A, Artifact A21)
- Prototyping_reformat/DatasetAnalysis/LIMUC/1_frozen_encoders/vit_frozen_logreg.ipynb (Appendix A, Artifact A21)
- Prototyping_reformat/DatasetAnalysis/LIMUC/1_frozen_encoders/clip_linear_baseline.ipynb (Appendix A, Artifact A21)

The fine-tuned baselines are implemented in:

- Prototyping_reformat/DatasetAnalysis/LIMUC/2_supervised_finetuning/finetune_resnet50.ipynb (Appendix A, Artifact A22)
- Prototyping_reformat/DatasetAnalysis/LIMUC/2_supervised_finetuning/finetune_vit_or_swin.ipynb (Appendix A, Artifact A22)

These baselines are not just included for completeness. They provide the lowest bar that a generative method must clear to be considered. In many cases, in medical imaging, a generative system may be attractive due to its human-friendly output, but still not achieve better performance than a simpler classifier on the variable of interest. By laying a foundation that first focuses on supervised performance, the proposed method

is evaluated against the best alternative that is currently justified rather than a naive baseline.

4.3.3 Zero-Shot Generative Baseline

In the third layer, a zero-shot generative severity baseline is introduced using the notebook listed in Appendix A, Artifact A8. This layer is crucial because it shows the unbridged transfer gap between multimodal fluency and constrained severity.

The zero-shot model uses a static prompt for the severity question and constrains the text outputs to be SCORE: X where X is in {0,1,2,3}. A rigid parser then extracts the severity score and explicitly flags noncompliant scores. This has two purposes. First, it maximises the performance of the zero-shot model in a well-defined format. Second, it ensures that the evaluation doesn't credit plausible responses that don't map to the label space.

Zero-shot evaluation is thus used as a diagnostic, rather than the suggested solution. It provides insight into what can and cannot be achieved from prompt-only transfer prior to in-domain adaptation.

4.3.4 Parameter-Efficient Generative Adaptation with LoRA

The main methodological innovation of this chapter is the parameter-efficient adaptation step implemented in the LoRA fine-tuning notebook listed in Appendix A, Artifact A9. This step fine-tunes a vision-language generation stack through Low-Rank Adaptation (LoRA) rather than full-blown end-to-end training (Hu et al. 2021).

LoRA is an appropriate approach for both methodological and scientific reasons. Pragmatically, it makes adapting a large multimodal model less memory- and compute-intensive. Scientifically, it enables the chapter to experiment with turning a large vision-language pretrained model into a clinically useful model for severity grading by adapting it for the task in contrast to free generation. The final architecture compared is BLIP2-Flan-T5-XL with LoRA adapters for the generative stack (J. Li et al. 2023; Hu et al. 2021; Appendix A, Artifact A9).

This design choice is in line with recent multimodal medical literature that stresses the importance of combining vision and language foundation models under task-aware constraints for clinical utility (Fiaidhi et al. 2022; Fiaidhi et al. 2023).

Two design choices are especially important. First, the target is constrained, so that the model generates the severity token under a fixed prefix rather than free text. Second, the objective is targeted to the label token, which eliminates the incentive to generate the decorative filler that is not relevant for the Mayo decision. That is, the model remains generative, but it is focused on a constrained clinical outcome.

4.3.5 Retrieval-Supported Extension Path

The repository includes a retrieval-supported design pattern in the Kvasir-VQA-x1 RAG-BLIP2 evaluation notebook listed in Appendix A, Artifact A23. This pattern is included in the dissertation because it is a harbinger of evidence-supported multimodal reasoning. But it is not part of the main argument in Chapter 4.

This is by design. If evidence support were provided simultaneously with a severity adjustment, we would find it more difficult to evaluate the effect of the latter. So, Chapter 4 is first designed to answer the question: can a generative model of severity surpass a strong supervised baseline on the internal task when the output space is constrained? After this question is answered, Chapter 5 addresses the separate problem of evidence-grounded query support.

4.4 Experimental Design and Evaluation Protocol

The design of the experiment aims to separate real task learning from biases due to the output mode. Being a mixture of classification and generation, the evaluation protocol should evaluate both prediction accuracy and controllability of the output.

4.4.1 Task Formulation and Output Modes

The system takes as input a single colonoscopy frame and a fixed prompt (e.g. "how severe is the disease"). The output is a Mayo score in the closed set {0,1,2,3}. While the main focus of the dissertation is on enhanced interaction with physicians, the task in Chapter 4 is still the Mayo score prediction. An optional evidence phrase is therefore considered an afterthought.

There are two evaluation lanes for the LoRA-adapted system:

- mode1 (lora_mode1_train): the model generates text under the constrained prompt format, after which a strict parser extracts SCORE: <0|1|2|3>.

- mode2 (lora_mode2_eval): the score is selected by candidate-label likelihood after the SCORE: prefix using a sequence_logprob strategy, without relying on free-text parsing.

These two lanes have distinct purposes. mode1 is the main generative lane because it retains a natural generation process while imposing explicit control on the output. mode2 is a pathological ablation of mode1: it eliminates the free-text parsing and reduces the task to selecting the score according to label probabilities. If mode2 performs as well or better than mode1, that would indicate that free-text generation isn't essential for the task. If it does not, then this indicates that the benefit arises from the adapted generation rather than the simple probability label shortcut.

4.4.2 Metric Bundle and Clinical Interpretation

No single metric is sufficient for this task. Chapter 4 therefore evaluates each lane using a set of complementary measures.

Metric	**Role in evaluation**
Accuracy	Overall correctness across all test cases
Macro-F1	Class-balanced discrimination under label imbalance
Balanced accuracy	Mean recall across classes, reducing majority-class dominance
QWK	Primary ordinal-agreement metric for Mayo 0-3 scoring
MAE / RMSE	Magnitude of ordinal error, penalizing distant mistakes
Per-class precision, recall, and F1	Class-specific error interpretation
Parse rate	Validity of generative outputs under strict extraction
Remission-oriented slice (0-1 vs 2-3)	Clinically simplified threshold behavior

Table 4.4. Metric Bundle

QWK is considered the primary metric because it is sensitive to correct ranked agreement, and penalises disagreements by large margins more than disagreements by small margins. It is therefore more suitable for the severity task than accuracy. We also report macro-F1 and balanced accuracy because the dataset is imbalanced and a model that overpredicts the majority classes could appear to perform well in terms of accuracy, but fail on the clinically relevant minority severities. Parse rate is reported so that the generative system is only trusted if its text can be reliably translated back to a score.

4.4.3 Reporting Policy, Statistical Checks, and Multi-Seed Aggregation

This chapter reports multi-seed aggregates rather than a single run. This is particularly important for the generative lane where the variance in the optimization could give a false sense of security. The supervised family is reported with seeds 11/23/42 and the final generative family with seeds 11/23/77.

We report the full uncertainty around aggregate metrics (e.g. QWK) so that the reader can assess whether reported improvements are statistically significant or just within the noise. If paired predictions are produced, paired tests (e.g. McNemar's test) are also reported. The chapter also contains a seed-level quality-control table for the mode1 generative lane, which shows that all runs were stable (no degeneracy) and produced predictions for all four classes.

This reporting practice follows the evidential standard in the dissertation. The purpose is not to report the highest number under variable conditions, but rather to ascertain whether the proposed modeling approach provides a consistent, stable, and interpretable improvement over the supervised baseline.

4.4.4 Final Training Configuration Used for Reporting

The official configurations in this chapter are collated from the persisted LIMUC reporting artifacts listed in Appendix A, Artifact A3, namely the multi-seed summaries for Pass 5 and Pass 6.

Lane	Core setup
Pass 5 supervised	ResNet50 fine-tuning, seeds 11/23/42, 15 epochs, batch size 16, learning rate 3e-4, weight decay 1e-4
Pass 6 generative	BLIP2-Flan-T5-XL with LoRA, seeds 11/23/77, 2 epochs, batch size 2, gradient accumulation 4, learning rate 5e-5, LoRA r=8, alpha=16, dropout 0.1, balanced sampling, label-token-only objective

Table 4.5. Final Configuration Summary for Reported Results

The inequality between supervised and generative setups is not an issue. The two families of models use different model structures and optimization constraints. It is not necessary for them to share the same hyper-parameters, but it is important that each family is trained under a sensible and consistent regime, and that the final comparison is between the multi-seed aggregates, rather than the few cherry-picked cases.

4.5 Results

The values cited in this section are from the persisted LIMUC reporting results and synchronized Chapter 4 figure assets listed in Appendix A, Artifacts A3 and A24.The analysis in this chapter is based on these fixed values.

4.5.1 Internal Multi-Seed Comparison on LIMUC

The main finding of this chapter is that the generative lane adapted to LoRA (mode1) achieves better results on the internal LIMUC test set on the headline metrics of particular interest for this task.

Lane	Seeds	Accuracy	Macro-F1	Balanced accuracy	QWK	95% CI (QWK)	Parse rate
Pass 5 supervised	11/23/42	0.737643	0.667330	0.670907	0.818649	[0.807920, 0.830582]	--
Pass 6 lora_mode1_train	11/23/77	0.781930	0.727920	0.736292	0.863656	[0.862382, 0.865836]	1.000000
Pass 6 lora_mode2_eval	11/23/77	0.548636	0.177135	0.250000	0.000000	[0.000000, 0.000000]	1.000000

Table 4.6. Internal Multi-Seed Results on LIMUC

Relative to the supervised baseline, mode1 improves internal QWK by +0.045007, macro-F1 by +0.060590, balanced accuracy by +0.065385, and accuracy by +0.044286. It's doing this while preserving a parse rate of 1.000000, which means that all of its generated answers (reported for the internal test set) can be translated into a score. This is significant because it indicates our proposed method does not gain in

scoring by producing unstable text. On the contrary, it achieves improved ordinal agreement and perfect format compliance under the constrained protocol.

The small QWK confidence interval for mode1 also points to stable behavior with respect to different seeds. This assessment is confirmed by the seed-level quality-control report, which indicates that all three runs reported pass the non-degeneracy tests and that each run is predicting all four classes. In contrast, mode2 is a poor model despite also having a parse rate of 1.000000. This immediate distinction gives us a clue that syntax is not sufficient; we also need semantics.

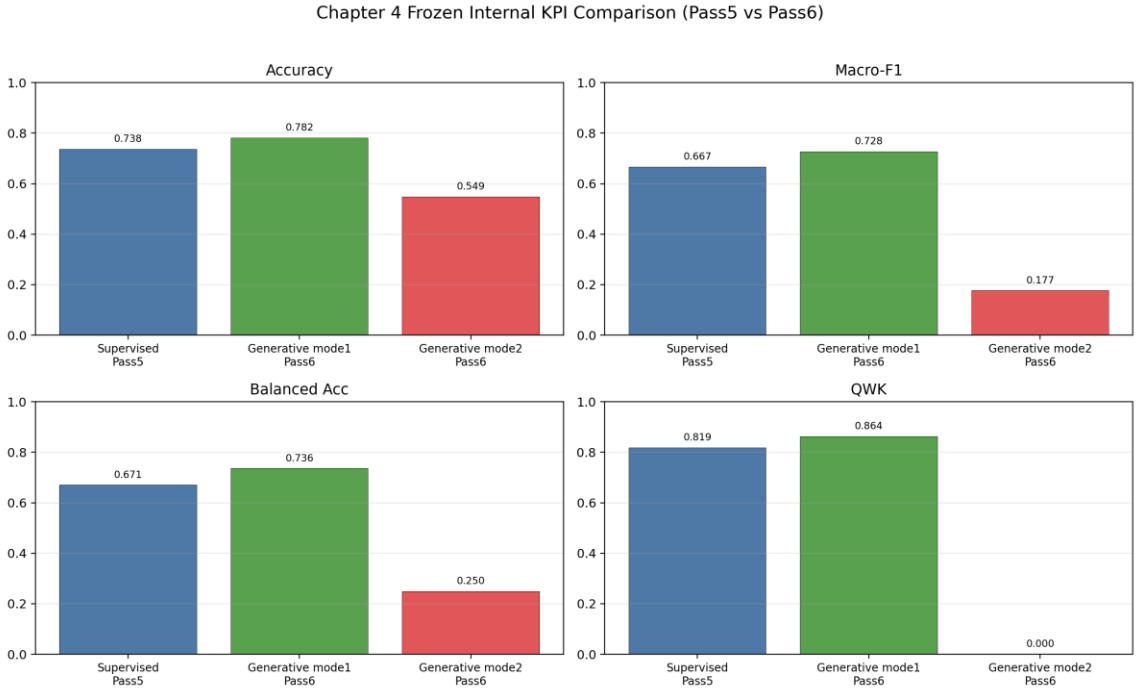

Figure 4.1. Internal LIMUC Metric Comparison

Figure 4.1. highlights the main internal finding: the LoRA-adapted mode1 lane outperforms the supervised baseline, whereas the mode2 ablation collapses despite perfect parse compliance (Appendix A, Artifact A24).

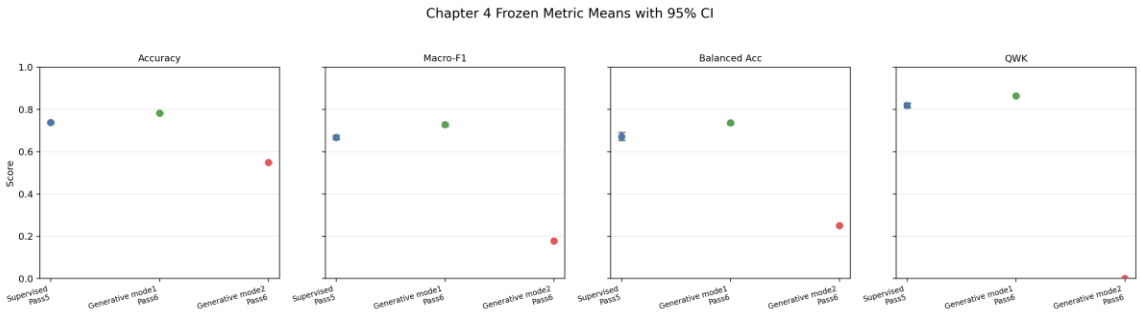

Figure 4.2. Multi-Metric Performance Profile for the Severity Lanes

Figure 4.2 shows that the advantage of mode1 is not confined to a single metric. The improvement is consistent across accuracy, macro-F1, balanced accuracy, and QWK (Appendix A, Artifact A24).

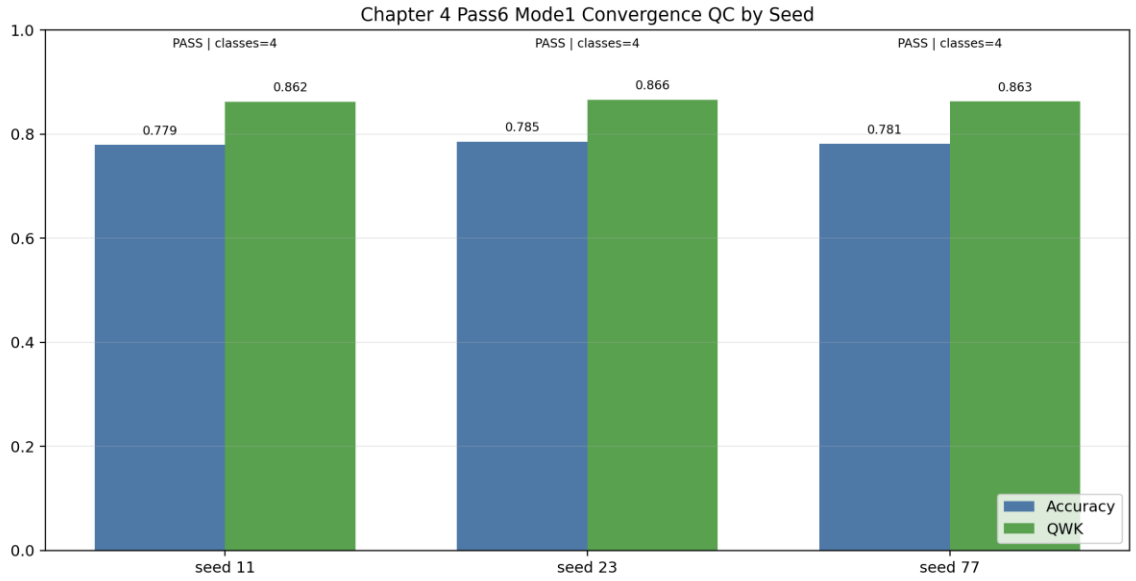

Figure 4.3. Seed-Level Quality-Control Profile for the Generative Lane

Figure 4.3. documents the stability of the reported generative lane across runs. The legacy filename is retained for reproducibility, but in this chapter the figure is used as a seed-level quality-control view rather than as the primary significance display (Appendix A, Artifact A24).

4.5.2 Class-Wise Behavior and Error Structure

Overall improvement is important but it does not reveal where the improvement occurs. The class-wise recall table shows that mode1 produces gains in recall for all Mayo classes, with larger gains in higher severity classes, which are more challenging to model.

Mayo class	Pass 5 supervised	Pass 6 mode1	Absolute gain
0	0.8043	0.8346	+0.0303
1	0.7011	0.7349	+0.0338
2	0.5782	0.7062	+0.1281
3	0.6000	0.6694	+0.0694

Table 4.7. Mean Per-Class Recall on LIMUC

The biggest improvement is seen in Mayo class 2, which goes from 0.5782 to 0.7062 in recall. This is a significant gain as class 2 represents a clinically relevant middle-to-high severity class and can be confused with nearby classes. There is also significant improvement in class 3, suggesting that the proposed generative adaptation does not merely focus on the most common remission or mild cases.

But the class-wise profile assures us that the task is still hard. The gain reduces, but does not remove, confusion at class boundaries, particularly between levels of severity. This makes sense given that the visual problem is difficult to categorise with subtle mucosal variations and frame-level ambiguity.

The mode2 ablation highlights the failure mode. It has a per-class recall of 1.0 for class 0 and 0.0 for classes 1-3, meaning that it simply predicts the majority class. This explains why mode2 can achieve 100% parse rate with a QWK of 0.0: it is generating valid labels, but the labels are not useful clinically because the ordinal class distinctions have collapsed.

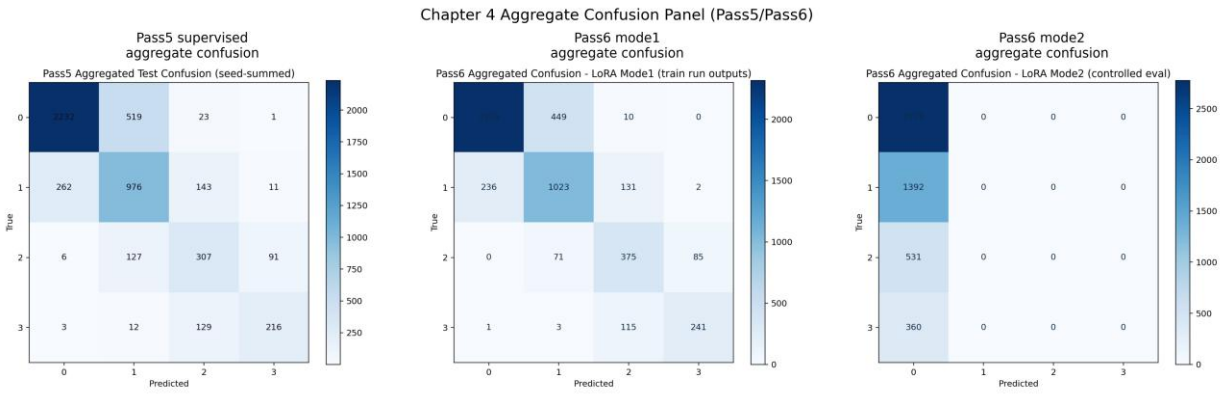

Figure 4.4. Confusion Matrix Comparison for Severity Prediction

Figure 4.4 visualizes the error structure underlying the aggregate metrics. The mode1 lane reduces confusion more evenly across classes, whereas mode2 degenerates into a class-0-dominant pattern (Appendix A, Artifact A24).

4.5.3 External Stress Test Under Domain Shift

The improvement observed on internal LIMUC does not, by itself, establish broader robustness. For that reason, the chapter includes an external-only HyperKvasir UC proxy evaluation as a stress test.

Lane	Internal QWK	External QWK	Delta (external-internal)	Internal parse rate	External parse rate
resnet50_supervised	0.828762	0.359597	-0.469165	--	--
vlm_lora_mode1	0.862752	0.000000	-0.862752	1.0	0.0
vlm_lora_mode2	0.000000	0.000000	0.000000	1.0	1.0

Table 4.8. Internal-to-External Performance Shift

The external performance is poor for all model types, but the mode of failure varies. The supervised neural baseline (ResNet50) performs far worse, but still achieves non-zero external agreement (QWK = 0.359597). In contrast, the mode1 generative lane breaks down entirely in this proxy scenario: external accuracy is 0.041128, macro-F1 is

0.019752, QWK is 0.0 and the parse rate is 0.0. This tells us that in the external proxy setting, the model is not just less accurate; it's no longer able to produce valid scores.

This is a negative result. It demonstrates that the success of the proposed generative lane is real, but not yet resilient to domain shift. The likely culprits are a combination of data shift, label-space mismatch caused by the proxy mapping, and style issues when the visual stimuli are no longer the same as the in-domain training distribution. The mode2 lane continues to collapse, internally and externally, which also confirms that it does not represent a possible alternative scoring strategy with the given design.

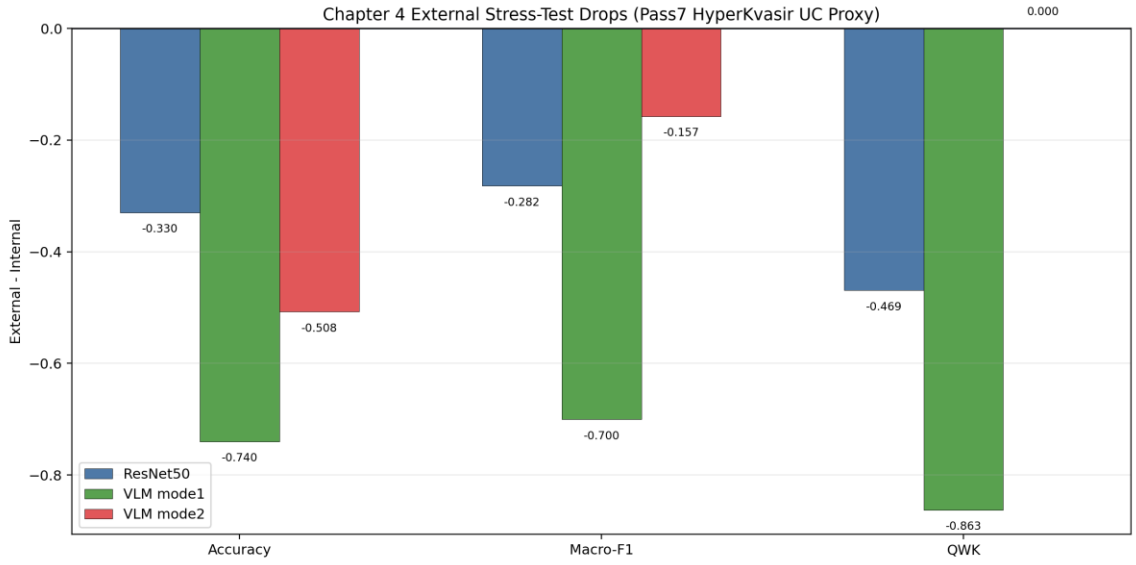

Figure 4.5. Internal-to-External Performance Shift Under Domain Shift

Figure 4.5 summarizes the extent of the external degradation. The figure should be interpreted as evidence of a robustness limitation rather than as a contradiction of the internal LIMUC improvement (Appendix A, Artifact A24).

4.6 Discussion

The findings support a particular claim: a vision-language model can be superior to a strong supervised baseline for internal severity grading of UC when its generation is tailored to the task via parameter-efficient training and output space design. This does not mean that open-ended multimodal generation is better than classification. Instead, it demonstrates that a constrained multimodal generation interface can be effective as an

ordinal scorer when the task, output space, and training objective are all in close alignment.

4.6.1 Why the Adapted Generative Lane Improves Internal Severity Grading

The mode1 lane is effective because it has three features that are usually considered independently in MedVQA. First, it is endowed with the representational power of a large pretrained vision-language model. Second, it is fine-tuned in-domain and not prompted. Third, it is constrained in its output enough that it is rewarded for issuing the decision token rather than for generating text that is more natural but less relevant to the task at hand.

This seems to be especially important for the minority and high-severity classes. The gains in class-wise recall suggest the LoRA-adapted model is not simply learning the majority labels better than the supervised model. Rather, it appears to find a better ordering-sensitive decision boundary for the four Mayo classes under the given prompt format. This is the main methodological finding of the chapter.

4.6.2 Why the Likelihood-Only Lane Fails

The success of mode2 is as interesting as the success of mode1. Likelihood-based label selection after a fixed prefix could have been a cleaner way to use the generative backbone as a classifier. It turns out to be equivalent to majority voting. The saved recall profile for each class shows that mode2 is predicting class 0 only. This is sufficient to preserve superficial syntactic validity, but not clinical usefulness.

The negative result is a simple example of one of the dissertation's messages: better control of the output is not necessarily better if it eliminates the mechanism of expression of the decision boundary learned by the adapted model. A generative system that is useful in clinical practice must be controlled, but it must be controlled in a way that allows task-relevant decision making. For the internal protocol, the mode1 control achieves this while mode2 control does not.

4.6.3 Implications for the Broader Dissertation

The main impact of Chapter 4 is methodological. It demonstrates that the road from benchmark MedVQA to multimodal systems for clinical use is not in open prompting. Rather, it lies in constrained task definition, explicit assessment of output validity, and the use of powerful supervised baselines. Chapter 4, in this way, is the link between the diagnostic benchmarking of Chapter 3 and the physician wrapper of Chapter 5.

The chapter also makes it clear what type of generative model should be taken forward. The severity component, proposed in Chapter 4, is not a chatbot. It is a controlled front-end component that is meant to translate an image into a stable severity signal according to a formatting prescription. It is this kind of component that can be safely integrated into a larger evidence-based decision support system.

4.7 Limitations and Claim Boundary

There are several limits to the results that must be discussed in the dissertation.

1. The task is frame-based rather than procedure-based. A single image can support severity estimation, but it is not equivalent to full-video or full-case clinical assessment.
2. Class-boundary ambiguity remains a persistent source of error, especially for adjacent Mayo categories such as 0 <-> 1 and 1 <-> 2.
3. The mode2 lane fails under the present design and should be treated as a negative-result ablation rather than as a viable alternative deployment mode.
4. The external HyperKvasir UC proxy evaluation uses mapped interval labels rather than native Mayo annotation. Its results are therefore informative for robustness analysis but insufficient for deployment claims.
5. Although mode1 improves internal QWK to 0.863656, the result remains below a 0.90 threshold and therefore still leaves meaningful headroom for future improvement.
6. Structured outputs that combine Mayo + evidence phrase are not part of the headline claim in this chapter. The contribution established here is limited to controlled severity scoring.

As such, the headline claim of Chapter 4 is focused and tight: under the fixed internal LIMUC protocol, the LoRA-adapted generative mode1 lane outperforms the official supervised baseline on ordinal and class-balanced severity metrics while ensuring 100% parse compliance. The chapter does not claim to be externally deployable, generally proficient at GI reasoning, or capable of strong severity transfer.

4.8 Chapter Summary and Transition to Chapter 5

In Chapter 4, the diagnostic results of Chapter 3 were turned into a proposed approach. It used LIMUC as the main empirical evidence source to describe the reproducible Mayo 0-3 severity task, set up a baseline with high supervised performance, and implemented a LoRA fine-tuned vision-language severity model, testing it under a multi-seed controlled experiment. The key empirical finding is this. The generative mode1 lane outperforms the supervised baseline on internal LIMUC in QWK, macro-F1, balanced accuracy, accuracy, and achieves perfect parse validity. At the same time, the external stress test confirms that this improvement should be considered in-domain rather than evidence of generalization.

This result delivers a component for the next phase of the dissertation. Chapter 5 uses the severity signal we produce here as a constant and embeds it into a doctor-facing shell that extends it with PICO extraction, retrieval, citation links and safety constraints. So, Chapter 4 is about the bounded severity-estimation problem while Chapter 5 is about how to make the bounded signal usable in a more traceable and actionable multimodal layer.

Chapter 5.

PICO-Grounded GenAI Wrapper for Physician Query Support

5.1 Chapter Purpose and Contribution

This chapter can be thought of as the integration chapter of the thesis. Chapters 1 and 2 set the conceptual framework for medical VQA and multimodal medical reasoning. Then, Chapter 3 established that reliability on GI tasks comes from task design rather than free generation, and Chapter 4 interpreted that finding into a reproducible UC severity module under a tight internal claim boundary. In this chapter, we take the frozen upstream module and couple it with an auditable, citational, and rule-constrained physician-query workflow.

The key innovation in this chapter is a shift in focus from model performance to workflow reliability. It is no longer just a question of whether a model will predict accurately on a test set, but whether an entire query-to-answer procedure will present evidence, uncertainty and policy constraints to a clinician for review.

This is in line with recent PICO-focused GenAI systems for evidence-based medicine and motivates query decomposition and response synthesis from sources rather than open-ended question answering (Mohammed and Fiaidhi 2024).

The contribution of this chapter is practical and system-level:

1. convert physician-style queries into structured PICO fields
2. retrieve evidence chunks conditioned on PICO intent
3. synthesize citation-linked claims under explicit safety rules
4. preserve compatibility with frozen Chapter 4 severity context through typed schemas
5. provide reproducible evaluation and completion-audit artifacts.

This chapter is therefore the intermediary between a controlled scoring engine (as in Chapter 4) and a safer interaction layer for doctors while retaining explicit constraints on the power of the system.

5.2 Boundary Conditions from Chapter 4 Freeze

For cross-chapter consistency, the upstream severity boundary used in this chapter is aligned with the frozen Pass 5/6/7 reporting policy from Chapter 4 rather than with exploratory runs. The official internal Chapter 4 reference points used here are:

Upstream lane (Chapter 4)	Accuracy	Macro-F1	Balanced accuracy	QWK	Parse rate
Pass 5 supervised	0.737643	0.667330	0.670907	0.818649	--
Pass 6 mode1 (primary generative lane)	0.781930	0.727920	0.736292	0.863656	1.000000
Pass 6 mode2 (negative ablation)	0.548636	0.177135	0.250000	0.000000	1.000000

Table 5.1. Chapter 4 Severity Reference Points Used in Chapter 5

In the Chapter 5 pass4 wrapper evaluation run, severity context ingestion is supported by schema (SeverityResult) but intentionally disabled for the reported benchmark pass (has_severity_context=false) to isolate wrapper behavior.

5.3 Design Objectives and Non-Objectives

5.3.1 Design objectives

1. enforce structured outputs with explicit claims and citations;
2. keep failure behavior visible through refusal or escalation rather than fabricated confidence
3. preserve deterministic local reproducibility without dependence on external APIs
4. keep module boundaries clear so the wrapper can be audited independently of Chapter 4 retraining.

5.3.2 Non-objectives in this chapter

1. no generation of patient-specific dosing recommendations
2. no claim of readiness for external clinical deployment
3. no replacement of physician judgment
4. no claim that the current internal knowledge-base coverage approximates guideline-complete retrieval

5.4 System Architecture and Contracts

5.4.1 End-to-end processing flow

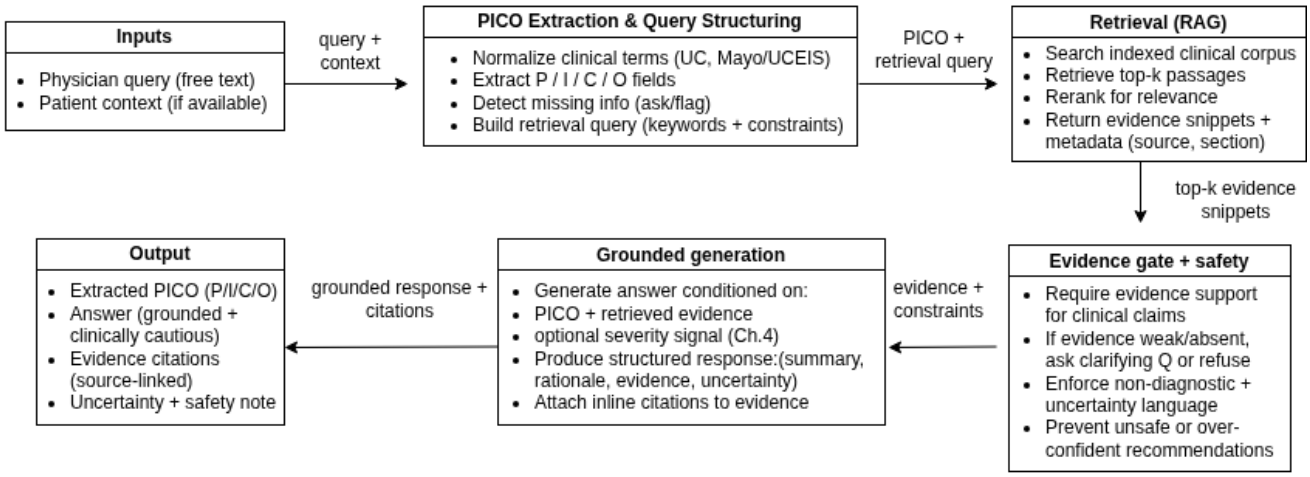

Figure 5.1. End-to-End Processing Flow of the PICO-Grounded Wrapper

5.4.2 Module-level implementation map

Module	Responsibility
schemas.py	Typed contracts (PicoFrame, SeverityResult, EvidenceChunk, Citation, WrapperInput, WrapperOutput)
pico_extract.py	Rule-based PICO extraction with optional LLM hook and safe fallback
kb_ingest.py	KB ingestion, chunking, and index metadata generation
retriever.py	Backend-aware retrieval with PICO-conditioned query composition
synthesis.py	Deterministic claim synthesis and citation attachment
safety.py	Refusal/escalation policy and disclaimer enforcement
wrapper.py	Full pipeline orchestration (extract -> retrieve -> synthesize)

ui_support.py	UI-side formatting and report helpers

Table 5.2. Chapter 5 Wrapper Module Responsibilities

5.4.3 Retrieval design used

From Prototyping_reformat/chapter5_pico_wrapper/results/kb_build_pass4_latest/kb_manifest.json:

- source files: 3
- documents: 3
- chunks: 12
- chunking: max_words=180, overlap_words=30, min_words=30, seed=42
- backend family available: keyword, tfidf, semantic_lsa, hybrid
- pass4 reporting backend: hybrid with reranker enabled.

5.4.4 Safety policy implementation

The wrapper applies explicit safety constraints. Dosing-sensitive prompts may trigger refusal behavior. Retrieval paths with limited supporting evidence trigger conservative uncertainty and limitation language. All outputs include clinician-review disclaimers. Claims excluded by policy are tracked separately during answer evaluation.

5.5 Experimental Protocol and Frozen Artifacts

5.5.1 Query and gold sets

From Prototyping_reformat/chapter5_pico_wrapper/data/queries/:

- queries.jsonl: n=50
- pico_gold.jsonl: n=20
- retrieval_gold.jsonl: n=10

5.5.2 Frozen pass used for chapter reporting

To remain consistent with the completion-audit record and the figure pipeline, this chapter reports the pass4 artifact set.

- KB: results/kb_build_pass4_latest/
- wrapper outputs: results/wrapper_eval_pass4_latest/
- wrapper output file: Prototyping_reformat/chapter5_pico_wrapper/results/wrapper_eval_pass4_latest/wrapper_outputs.jsonl
- evaluations: results/eval_pass4_latest/
- completion audit: results/chapter5_completion_audit_ch5_freeze_20260306/
- pipeline summary: results/pipeline_pass4_latest/pipeline_summary.json

Key wrapper config (results/wrapper_eval_pass4_latest/run_config.json):

Parameter	Value
run_id	chapter5_wrapper_20260305T024427Z_6nx5r2
mode_requested	baseline
n_queries	50
retrieval_k	5
retrieval_backend	hybrid
rerank_enabled	true
rerank_pool	20
rerank_alpha	0.2
min_top_score_for_answer	0.18
min_mean_score_for_answer	0.12
min_retrieved_for_answer	2
has_severity_context	false

Table 5.3. Pass 4 Wrapper Evaluation Configuration

5.6 Results

5.6.1 PICO extraction quality

Source: Prototyping_reformat/chapter5_pico_wrapper/results/eval_pass4_latest/pico_eval.json.

Field	Precision	Recall	F1
Population	1.0000	1.0000	1.0000
Intervention	0.6250	1.0000	0.7692
Comparator	0.8800	1.0000	0.9362
Outcomes	0.4000	0.5000	0.4444
Severity anchors	0.4667	1.0000	0.6364

Timeframe	1.0000	1.0000	1.0000
Setting	0.8000	0.5000	0.6154
Constraints	0.0000	0.0000	0.0000

Table 5.4. PICO Extraction Performance by Field

Aggregate values:

- required-field macro-F1 (P/I/C/O + severity_anchors): 0.7572 (n=20)
- all-field macro-F1: 0.6752

The extractor is designed to be recall-heavy for the core required fields, but its precision is weaker for outcomes and severity anchors, indicating broad lexical matching behavior.

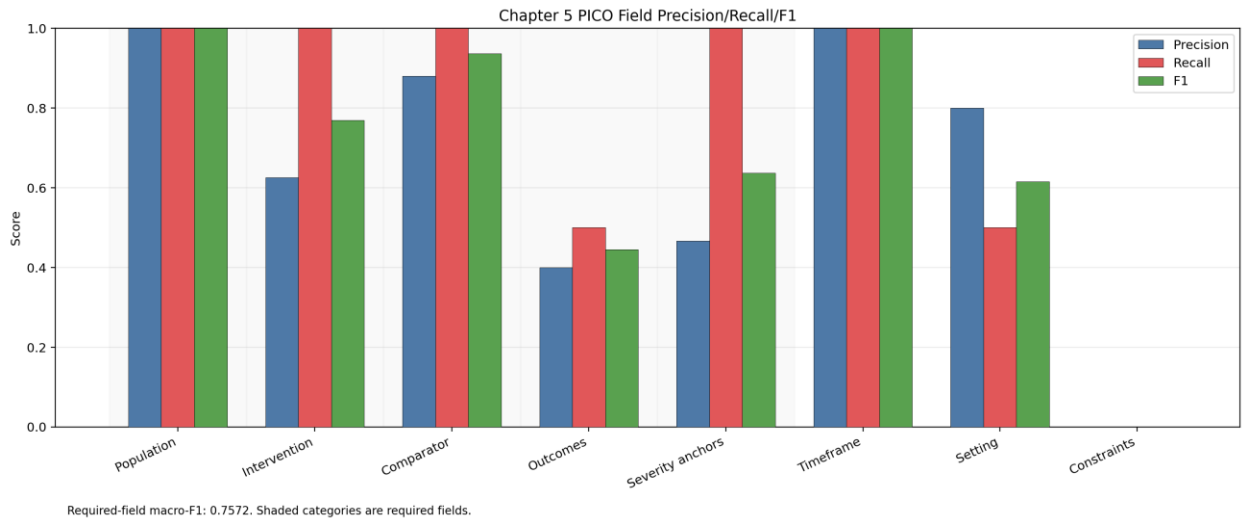

Figure 5.2. PICO Extraction Performance by Field

Figure 5.2 presents the per-field PICO extraction scores, with shaded groups indicating the required fields.

5.6.2 Retrieval Quality and Uncertainty

Source: Prototyping_reformat/chapter5_pico_wrapper/results/eval_pass4_latest/retrieval_eval.json.

Metric	@1	@3	@5
Precision@k	0.2000	0.1667	0.1600
Recall@k	0.1000	0.2500	0.4500
Hit rate@k	0.2000	0.3000	0.6000

Table 5.5. Retrieval Precision, Recall, and Hit Rate at k

Bootstrap 95% confidence intervals (2000 iterations, seed = 42) are reported for the retrieval metrics.

Metric	@1 CI	@3 CI	@5 CI
Precision@k	[0.00, 0.50]	[0.00, 0.3667]	[0.08, 0.26]
Recall@k	[0.00, 0.25]	[0.00, 0.55]	[0.20, 0.70]
Hit rate@k	[0.00, 0.50]	[0.00, 0.60]	[0.30, 0.90]

Table 5.6. Bootstrap Confidence Intervals for Retrieval Metrics

The results indicate that top-5 retrieval provides workable coverage on the current small knowledge base, although the wide intervals reflect the limited size of the retrieval-gold sample (n = 10).

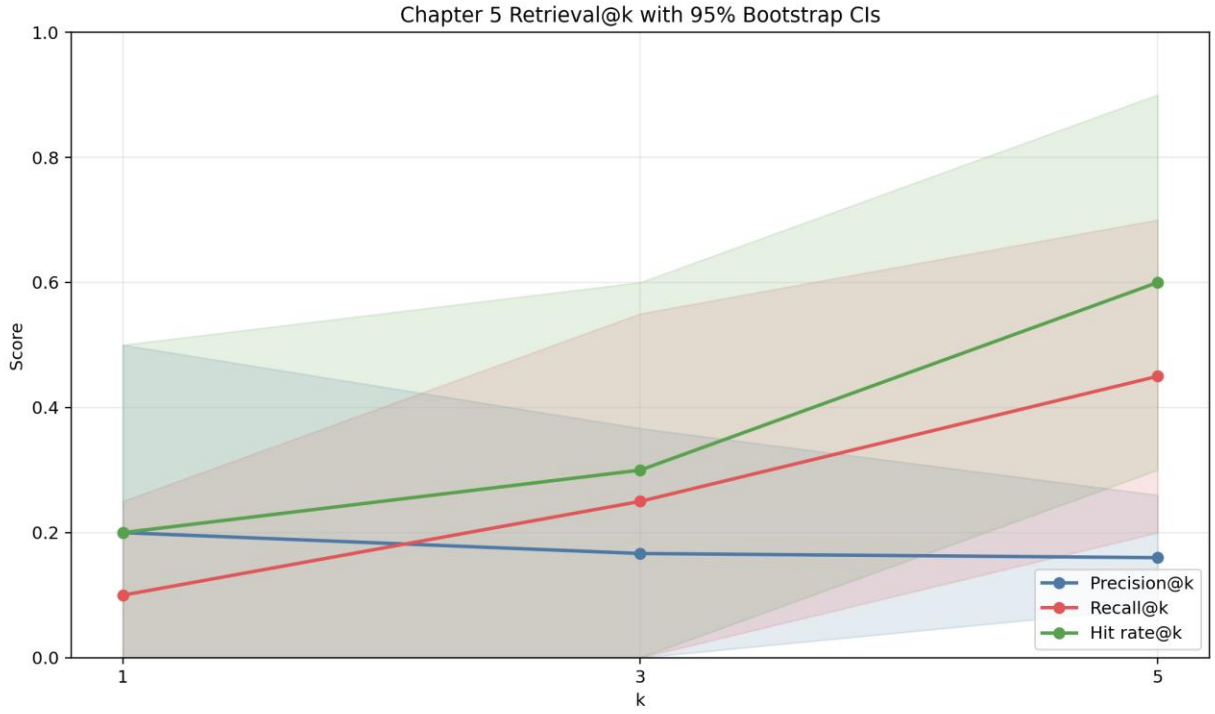

Figure 5.3. Retrieval Precision, Recall, and Hit-Rate Profiles with Confidence Intervals

5.6.3 Retrieval Ablation (Backend/Rerank Sensitivity)

Source: Prototyping_reformat/chapter5_pico_wrapper/results/eval_pass3_ablation/retrieval_ablation_summary.tsv.

Case	Backend	Rerank	Alpha	Hit@1	Hit@5	Recall@5
keyword_no_rerank	keyword	no	0.35	0.20	0.50	0.30
hybrid_rerank_a035	hybrid	yes	0.35	0.20	0.50	0.35
hybrid_rerank_a050	hybrid	yes	0.50	0.20	0.50	0.35
tfidf_no_rerank	tfidf	no	0.35	0.10	0.60	0.45
tfidf_rerank_a035	tfidf	yes	0.35	0.20	0.60	0.45
hybrid_no_rerank	hybrid	no	0.35	0.20	0.60	0.45
hybrid_rerank_a020	hybrid	yes	0.20	0.20	0.60	0.45

Table 5.7. Retrieval Backend and Reranking Ablation Results

The selected default in pass4 is hybrid + rerank alpha = 0.20, as this setting ties for the best Recall@5 while preserving stronger Hit@1 than non-reranked TF-IDF.

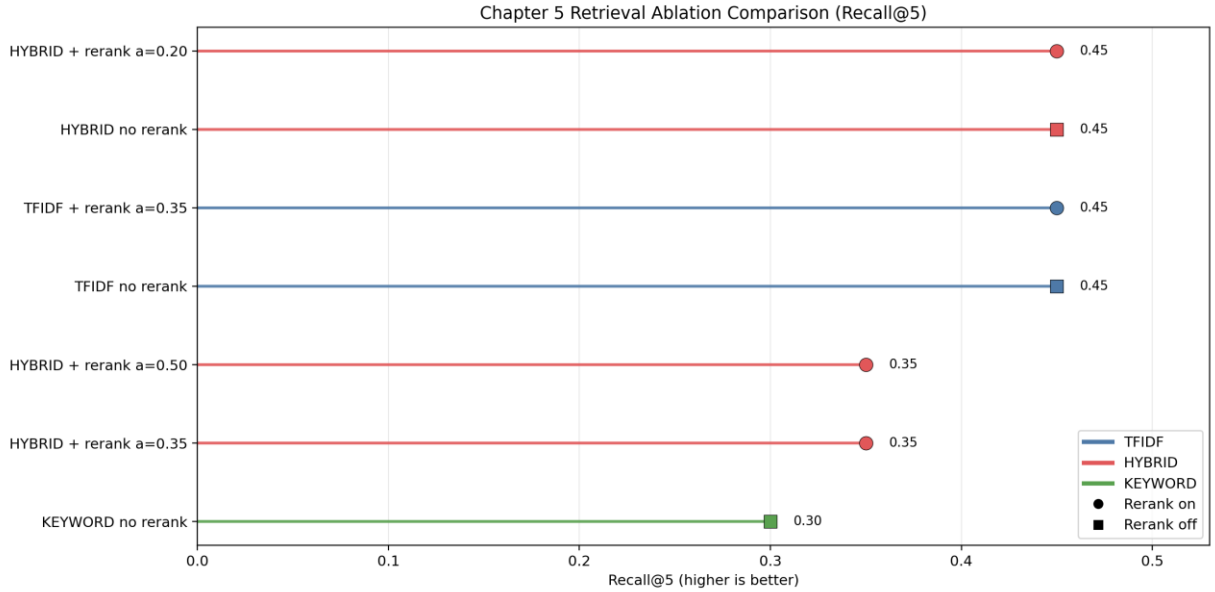

Figure 5.4. Retrieval Recall@5 Across Backend and Reranking Configurations

Figure 5.4. compares Recall@5 across backend and rerank configurations. Several settings tie at the top under this small benchmark, which suggests that the present comparison should be interpreted cautiously.

5.6.4 Answer quality and citation grounding

Source: Prototyping_reformat/chapter5_pico_wrapper/results/eval_pass4_latest/answer_eval.json.

Metric	Value
Outputs evaluated	50
Claims extracted	142
Claims evaluated	138
Policy claims excluded	4
Refusal count	4
Refusal rate	0.0800
Citation coverage	1.0000
Citation correctness (heuristic)	1.0000
Claim support (heuristic)	1.0000
Claim support (strict)	0.8696
Contradiction proxy	0.0000
Hallucination proxy	0.0000
Citation link integrity	1.0000

Table 5.8. Answer Quality and Citation-Grounding Metrics

Strict support is intentionally harder than the heuristic overlap checks (min_overlap_ratio=0.25, min_overlap_terms=3), providing a tighter baseline for future manual review.

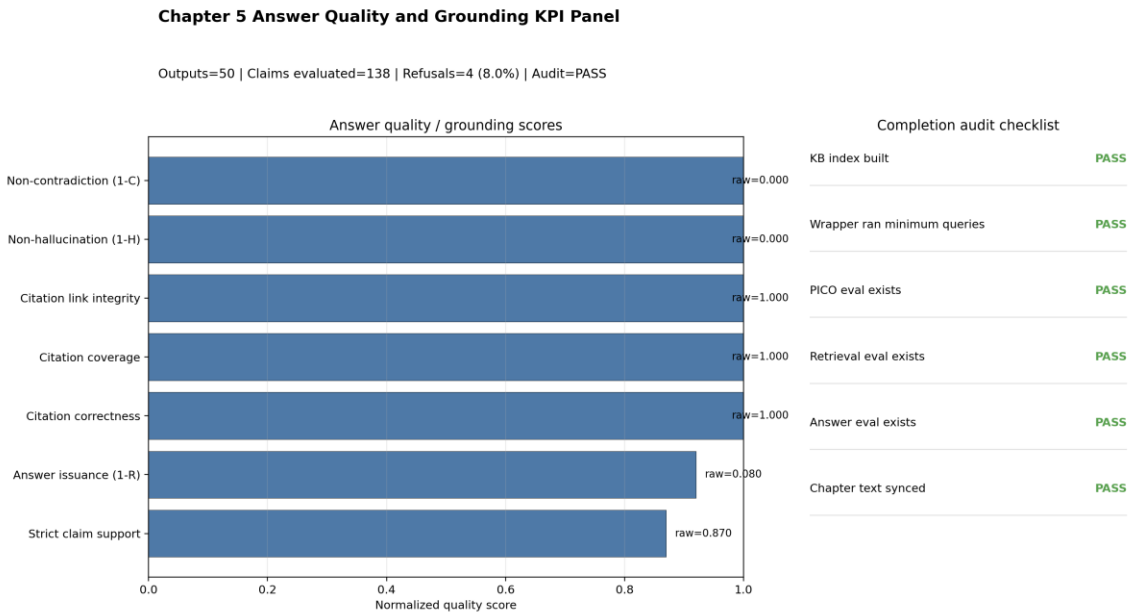

Figure 5.5. Answer Quality, Refusal, and Completion-Audit Summary

Figure 5.5 presents a composite panel showing the main answer-quality indicators, refusal rate, and completion-audit checklist status.

5.7 Error Analysis and Observed Failure Modes

The pass4 evidence suggests a number of recurrent problems.

1. **Outcome extraction over-triggering:** precision for outcomes (0.4000) is much lower than recall (0.5000), suggesting over-coverage based on word matching.
2. **Severity-anchor over-triggering:** precision for severity_anchors (0.4667) is lower than recall (1.0000), again showing wide term matching.
3. **Top-rank retrieval fragility:** recall improves as k increases, but top-1 behavior remains brittle (Recall@1 = 0.1000).
4. **Evidence-threshold sensitivity:** strict claim-support scoring (0.8696) is lower than heuristic support (1.0000), showing dependence on overlap thresholds.

5. **Policy handling effects:** refusals are necessary for safety, but they reduce the answer-issuance rate (92% non-refusal issuance).

These effects are expected in a baseline that prioritizes traceability and deterministic behavior over free-form fluency.

5.8 Threats to Validity and Limitations

1. Limited knowledge base: the knowledge base is limited to three documents and 12 chunks.

2. Small PICO and retrieval gold sub-sets: The retrieval gold (n = 10) and the PICO gold (n = 20) sub-sets are helpful for initial verification because they establish the wrapper, extraction, retrieval, citation, and refusal procedures work in a consistent manner. But they are too small for generalisation. But future research should increase the size of these gold sets with larger, clinician-adjudicated query banks, multiple annotators, inter-rater agreement reporting and more diverse query types, diseases, interventions, outcomes and clinical contexts. This would enable the wrapper to be evaluated not just as a proof-of-concept device, but also as a more generalised evidence-based decision support system.

3. Heuristic grounding metrics: citation correctness and hallucination proxies are lexical in nature and therefore need to be complemented by clinician semantic review.

4. No external-API LLM dependency in the baseline: this improves reproducibility, but it also means the current system does not yet benchmark richer model-assisted synthesis.

5. No deployment claim: this chapter demonstrates engineering feasibility and reproducibility, not bedside readiness.

5.9 Reproducibility and Completion Audit Status

Completion audit source:

- Prototyping_reformat/chapter5_pico_wrapper/results/chapter5_completion_audit_ch5_freeze_20260306/chapter5_completion_report.json

- Prototyping_reformat/chapter5_pico_wrapper/results/chapter5_completion_audit_ch5_freeze_20260306/chapter5_completion_report.md

Audit status: PASS (6/6 checklist items passed), including artifact existence, minimum query coverage, and chapter-text synchronization checks.

Core scripts:

- scripts/build_kb.py
- scripts/make_queryset.py
- scripts/run_wrapper.py
- scripts/run_ui.py
- scripts/eval_pico.py
- scripts/eval_retrieval.py
- scripts/eval_answers.py
- scripts/chapter5_completion_audit.py
- scripts/run_full_pipeline.py

5.10 Chapter 5 Claim Guardrail

Allowed headline claims:

1. Chapter 5 offers a stable, referenceable, citation-aware implementation of the locked-down upstream severity capability from Chapter 4.
2. Mandatory-field PICO extraction and top-5 retrieval have acceptable baseline behavior on the internal benchmark artifacts.
3. Citation-based synthesis and refusal are working in pass4 and can be inspected in the persisted results.

Disallowed headline claims:

1. The internal knowledge base (as of now) and the synthetic or small gold subsets do not warrant any claim of readiness for external clinical use.
2. Retrieval and grounding should not be touted as replacements for semantic adjudication by clinicians.

3. This chapter should not be interpreted as replacing the severity-model validation established in Chapter 4.

5.11 Summary

This chapter provides a full PICO-based wrapper layer that is structured, safe, audit-guided and reproducible. The wrapper does not mask uncertainty through citation, enables refusal, and produces clean audit trails of artifacts.

This chapter offers a modular alternative to an end-to-end generator, by freezing the visual evidence of severity from Chapter 4 together with explicit control of retrieval-grounded generation. Future work should build on the current baseline in terms of greater size and curation of the evidence store, better supervision of the retrieval process, and studies with clinician-scored semantic grounding, while maintaining the same claim discipline and auditability standards as the rest of the dissertation.

Thus, the limited PICO and retrieval evaluation should be considered as feasibility evidence rather than general clinical evidence. The same approach could be applied to other conditions or specialties because PICO extraction, retrieval, citation-linking, refusal and audit logging are not specific to ulcerative colitis. But disease-specific evidence sources would be needed for other conditions, together with clinician-reviewed gold standards, and possibly different query templates and external validation on larger data sets to support a general clinical claim.

Chapter 6.

Conclusions and Future Research

6.1 Chapter Purpose and Position in the Dissertation

This chapter wraps up the dissertation with a synthesis of the technical, empirical and translational work in Chapters 1 to 5. Our work started with a practical challenge in clinical AI: while GI-endoscopy visual question answering (VQA) systems may perform well on benchmark tests, their clinical value relies on accurate performance under class imbalance, robustness under domain shift, conveying uncertainty, and making evidence-based responses.

The dissertation tackled this issue through a multi-step work-flow, rather than focusing solely on model optimisation:

1. clarify the clinical problem, scope, research questions (Chapter 1)
2. define the MedVQA and GI-endoscopy literature and benchmarking state-of-the-art (Chapter 2)
3. empirically audit existing model families using persisted GI artifacts (Chapter 3)
4. develop a controlled generative UC severity module with frozen reporting boundaries (Chapter 4)
5. integrate a severity-compatible, PICO-grounded, citation-aware physician query tool (Chapter 5).

The first conclusion is that significant improvements in clinical GI MedVQA rely on constrained, auditable systems engineering and constrained claims, not unconstrained generation.

The second, equally important conclusion is that the dissertation produces not just a model result, but also a deployment logic. Hence, it doesn't just deliver the score table. Rather, it provides a decision process that clarifies when the evidence is sufficient to support limited clinical decision support, and when a decision should be escalated,

delayed, or denied. That is important in medical AI, where it is better to underestimate generalization than to overestimate it.

6.2 Consolidated Narrative of What Was Demonstrated

Across the full pipeline, five cross-chapter findings are consistent and mutually reinforcing.

1. Data structure and answer space control reliability. GI MedVQA resources vary widely in question families, answer cardinality, supervision method and label semantics, so aggregate performance measures can mask clinically significant flaws (Simula Datasets n.d.; Simula 2024; Simula n.d.; Smedsrud et al. 2021).

2. Constrained and supervised pipelines continue to be a reliability benchmark. As Chapter 3 demonstrated, naive open-ended generation is unreliable in the face of GI-specific constraints, particularly underrepresented classes and output variability.

3. Constrained generative adaptation can outperform high-quality supervised systems. In the internal evaluation in Chapter 4, the frozen generative model lane achieved higher QWK, macro-F1, balanced accuracy, and accuracy without sacrificing parse compliance (Appendix A, Artifacts A3 and A24).

4. Domain-shift robustness is not solved by in-domain gains alone. The external proxy evaluation in Chapter 4 showed substantial degradation, which means that the internal improvements must be interpreted as evidence of bounded internal validity rather than of universal transfer performance.

5. Workflow-level safeguards are essential for physician-facing use. Chapter 5 showed that citation linkage, refusal and escalation behavior, uncertainty surfacing, and completion-audit traceability can be operationalized within a reproducible wrapper (Appendix A, Artifacts A10–A15 and A17–A20).

These results, therefore, support a practical translational stance: reliability and safeguards should first be improved, and the capabilities of a model should only be grown under governance.

The findings also show a clear relationship among chapters. Chapter 2 defines what should be measured, Chapter 3 shows that current approaches are insufficient, Chapter 4 explores whether controlled generation can improve severity reliability, and

Chapter 5 explores whether that improved modeling can be delivered to physician use without removing safeguards.

6.3 Integrated Answers to Research Questions

Chapter 1 defined six research questions (RQ1-RQ6). Their final evidence-backed answers are summarized below.

Research Question	Final Answer from Dissertation Evidence
RQ1 (Coverage): What clinically relevant GI question families and answer spaces are represented in available datasets?	Coverage exists, but it is uneven. Dominant question families, such as yes/no and template-based prompts, are overrepresented relative to complex reasoning tasks and high-risk edge-case queries. Unless this imbalance is controlled, it can skew evaluation and create a risk of overestimation (Simula Datasets n.d.; Simula 2024; Simula n.d.; Smedsrud et al. 2021).
RQ2 (Comparative reliability): Are constrained/discriminative approaches more reliable than naive zero-shot open generation for GI MedVQA?	Yes, in current repository evidence and task settings. Chapter 3 shows constrained/supervised behavior as a recurring reliability anchor across GI datasets.
RQ3 (Failure modes): Which failure classes dominate current systems?	Class imbalance sensitivity, output-space mismatch, mapping/OOV instability, and domain-shift fragility emerged as dominant recurring failures.
RQ4 (Severity robustness): How reliably can models handle UC severity-oriented VQA, including severe classes?	Controlled adaptation in Chapter 4 materially improved ordinal and class-balanced reliability internally, but minority-severity challenges and external robustness limitations remain (Appendix A, Artifacts A3, A9, and A24).
RQ5 (Clinical output format): Which output style better supports clinician trust and usability?	Structured outputs with citation linkage, uncertainty, and policy-bounded limitations are more defensible for clinician review than unconstrained fluent responses under current evidence (Appendix A, Artifacts A10–A12 and A17–A20).
RQ6 (Evidence-aware extension): Can retrieval-grounded reasoning be integrated without compromising core visual-grounded behavior?	Yes, but only as a modular extension under explicit boundaries. Chapter 5 demonstrates a reproducible PICO-grounded wrapper with auditability and safety controls while preserving a constrained claim scope (Appendix A, Artifacts A10–A20).

Table 6.1. Integrated Answers to the Research Questions

6.3.1 Interpretation of RQ Closure

The RQ closure is intentionally bounded. The dissertation provides stronger evidence for method reliability and pipeline feasibility than for clinical deployment readi-

ness. It does not claim an official state-of-the-art or leaderboard result on ImageCLEFmed MEDVQA-GI 2025. Instead, validation is limited to persisted local evaluations: ImageCLEF MEDVQA-GI 2023 closed-label validation, LIMUC Mayo 0-3 severity evaluation, external-proxy stress testing, and Chapter 5 PICO/retrieval wrapper audits. These results support research-grade feasibility, not broad clinical deployment.

6.3.2 What RQ Closure Means for Clinical Translation

In a translation sense, the most valuable result of RQ closure is not the greatest metric improvement, but rather a specific evidence level attached to each claim. In other words, the final RQ closure supports three claims

1. the system is ready for research decision-support experiments with supervision
2. the system is not yet suitable for unsupervised clinical autonomy
3. future deployment is more dependent on the evidence of robustness and governance rather than relative gains in in-domain benchmarks.

This approach preserves the alignment of the dissertation to medical safety and the value of the technical achievements that have been made.

6.3.3 Examiner-Facing RQ Synthesis Table

To aid readability during the defence or viva, Table 6.2 maps each research question to the chapter evidence that supports an answer and to the final, bounded contribution claim supported by the evidence.

Research question	Primary chapter(s) answering it	Key evidence used	Bounded final claim
RQ1 (coverage of GI MedVQA tasks and answer spaces)	Chapter 2, Chapter 3	GI dataset/task-family mapping and benchmark audit (Simula Datasets n.d.; Simula 2024; Simula n.d.; Smedsrud et al. 2021)	Coverage is present but uneven; evaluation must control for skewed question/answer distributions.
RQ2 (constrained vs naive open generation reliability)	Chapter 3, Chapter 4	Cross-family reliability comparisons and LIMUC internal anchor (Appendix A, Artifacts A1–A5)	Constrained/supervised pipelines remain a necessary reliability baseline

			under current GI settings.
RQ3 (dominant failure modes)	Chapter 3	Failure taxonomy across imbalance, lexical mapping, and OOV behavior	Reliability failures are systematic, not random, and should drive architecture constraints.
RQ4 (UC severity robustness, including severe classes)	Chapter 4	Internal multi-seed LIMUC metrics plus class-wise analysis (Appendix A, Artifacts A3, A9, and A24)	Controlled generative adaptation improves internal ordinal and balanced performance, but does not solve external robustness.
RQ5 (clinician-facing output form)	Chapter 5	Wrapper output audits for citations, uncertainty, and refusal behavior (Appendix A, Artifacts A10–A12 and A17–A20)	Structured, citation-aware outputs are more defensible than unconstrained free-form responses.
RQ6 (evidence-grounded extension feasibility)	Chapter 5	PICO extraction, retrieval, and synthesis pipeline with typed contracts (Appendix A, Artifacts A10–A20)	Retrieval-grounded extension is feasible as a modular layer when bounded by explicit policy and audit controls.

Table 6.2. Research Question Closure Matrix

6.4 Final Contributions

The contributions of this dissertation can be broken into methodological, empirical and engineering contributions. They collectively define a reproducible design pattern for building multimodal medical AI systems: define the boundaries first, then optimize the performance inside the boundaries, and finally use the boundaries to explicitly control system behavior.

6.4.1 Methodological contributions

1. A reproducibility-first, claim-bounded evaluation strategy applied across all core chapters.

2. A staged architectural logic that separates severity-scoring reliability from physician-query reasoning reliability.

3. A concrete integration framework in which controlled generation is paired with structured evidence presentation rather than with a fluency-first free-text design.

This research approach reduces hidden variables between model quality and interface design. In the absence of this decoupling, we can hide model-level weaknesses and shortfalls with wrapper-level improvements, and vice versa.

6.4.2 Empirical contributions

1. A GI MedVQA map (Chapter 3) of reliable strengths and weaknesses across dataset regimes.
2. A frozen internal UC severity package (Chapter 4) where controlled generative adaptation surpassed a strong supervised baseline on primary ordinal and balanced metrics (Appendix A, Artifacts A3, A9, and A24).
3. A frozen PICO-driven wrapper benchmark (Chapter 5) in which citation-linked synthesis, refusal-aware response, and artifact-auditable generation were exhibited (Appendix A, Artifacts A10–A20).

So the empirical contribution is additive. Chapter 3 sets the baseline risk, Chapter 4 shows limited improvement, and Chapter 5 demonstrates evidence-grounded interaction subject to policy constraints.

6.4.3 Engineering contributions

1. End-to-end traceability from data preparation to chapter-level artifact synchronization.
2. Typed interfaces and persisted JSON/JSONL contracts that support wrapper reproducibility and auditability (Appendix A, Artifacts A10–A13 and A17–A20).
3. Completion-audit gating that enforces consistency between chapter text and artifacts at freeze time (Appendix A, Artifact A14).

These engineering advances are important because reproducibility is a functional requirement for medical AI. The dissertation artifacts are set up so that experiments can be reproduced, results can be inspected, and claims can be audited by others.

6.5 Revisit of Dissertation Hypotheses

The working hypotheses introduced in Chapter 1 can now be revisited.

1. H1: Constrained and supervised approaches are more reliable than naive zero-shot free generation. The results in Chapter 3 indicate this as do the ablation effects in Chapter 4.

2. H2: Severe minority classes are a major bottleneck. This is also supported. While Chapter 4 delivered improvements, these are not without severe class and boundary risks.

3. H3: Evidence-augmented extensions can enhance interpretability, without reducing the core visual grounding, when suitably constrained. This is supported as a baseline finding by the Chapter 5 wrapper results, within the acknowledged limitations (Appendix A, Artifacts A10–A20).

Taken together, these outcomes indicate that the central technical strategy of the dissertation is supported in directional terms.

However, the levels of support for the three hypotheses vary. H1 is strongly supported within the scope of tests. H2 is supported, but with minority-class risk. H3 is supported for the baseline wrapper, but is dependent on retrieval quality and policy settings. The nuanced interpretation is critical because it does not equate all of the hypothesis outcomes.

6.6 Practical Implications for Clinician-Facing AI

Given the current scope of the dissertation, three practical implications can be asserted.

1. Evidence and uncertainty must be treated as first-class outputs. Citation-linked claims accompanied by explicit uncertainty and limitations are safer than fluent responses that provide no evidential support.

2. Reliability must be assessed at both the model and workflow levels. A strong scoring component is not sufficient if downstream retrieval, synthesis, and safety behavior remain uncontrolled.

3. Negative findings are operationally informative. The mode2 collapse (Chapter 4) and the external shift are not disasters of the dissertation; they are limits that prevent overreaching and help direct next-step work.

Two further practice-level implications follow from the same evidence.

4. Escalation behavior should be understood as a feature rather than a defect. Refusal or defer-to-clinician outputs are necessary safety mechanisms when evidential confidence is weak.

5. Traceability should be considered prior to rollout pilots. Adding traceability after deployment is expensive and makes later claims to reliability less convincing.

6.7 Limitations and Validity Boundaries

While the advances presented in this dissertation are significant, the results are specifically limited by several constraints.

1. The main claims are based on internal frozen benchmark protocols, but not on prospective production systems.

2. The PICO and retrieval gold subsets in Chapter 5 are small and only suitable for benchmarking, not clinical deployment.

3. Grounding tests include heuristic elements and thus require larger-size semantic verification studies with clinicians.

4. Domain generalization is not addressed, especially in the presence of label-space and distribution shift.

5. The wrapper is a decision-support prototype and does not generate patient-specific treatment execution instructions.

6. To isolate the behavior of the wrapper, the Chapter 5 reporting freeze used has_severity_context=false; the cumulative effect of injecting severity context is an open question.

These limitations do not negate the contribution; they define the conditions under which the conclusions are valid.

6.7.1 Threats to Validity Framing

The above limitations can be expressed in terms of the threats to validity.

1. Internal validity threat: estimated benefits may be influenced by data set artifacts and assumptions in the frozen protocol.

2. External validity threat: generalization to new institutions, devices, patient populations, disease areas, and official challenge protocols remains unproven and requires multi-center, externally benchmarked validation.

3. Construct validity threat: the benchmark measures may not capture all the perceived utility and safety.

4. Conclusion validity threat: studies based on small subsets can increase variance and reduce confidence in fine-grained comparative claims.

These categories help to interpret the findings of the dissertation and identify what evidence is required for future studies.

6.8 Future Research Agenda

Future work should proceed in phased, testable increments. These directions follow prior work on clinically generated medical QA datasets, MedVQA benchmark construction, explainable and retrieval-grounded medical VQA, PICO-based evidence synthesis, and ulcerative-colitis severity validation (Lau et al. 2018; Ben Abacha et al. 2019; Lin et al. 2023; Mohammed and Fiaidhi 2024; Stidham et al. 2019; Takenaka et al. 2023).

6.8.1 Near-term research priorities

1. Enrich the Chapter 5 knowledge base with guideline, trial, and clinical-evidence anchors, following PICO-based evidence synthesis and retrieval-augmented medical VQA directions (Mohammed and Fiaidhi 2024; Karim and Uzuner 2025; Sial et al. 2025).

2. Expand the PICO, retrieval-relevance, evidence-span, and severity-anchor gold sets through independent annotation by at least two domain-informed annotators, followed by adjudication of disagreements and inter-annotator agreement reporting. For categorical labels, agreement should be reported with Cohen's kappa or Fleiss' kappa depending on the number of annotators; for ordinal severity labels, weighted kappa should be used; and for evidence-span marking, span-level overlap or F1 should be reported. This direction follows prior MedVQA and clinically generated QA dataset work that emphasizes curated question-answer resources and reliable expert annotation (Lau et al. 2018; Ben Abacha et al. 2019; Liu et al. 2021; Lin et al. 2023).

3. Increase precision of extraction for outcomes and severity anchors, without compromising recall.

4. Provide clinician-led semantic grounding review pipelines that move beyond lexical-overlap heuristics toward explainable and evidence-grounded MedVQA evaluation (Nguyen et al. 2025; Sial et al. 2025).

6.8.2 Mid-Term Robustness and Integration Priorities

1. Perform controlled wrapper studies with severity context (has_severity_context=true) and report the improvement.

2. Provide stronger retrieval supervision and ranking objectives using clinician relevance judgments.

3. Explore calibrated abstention that accounts for retrieval and severity uncertainty, as well as safety.

4. Extend the system from frame-level reasoning toward temporally coherent sequence-level or procedure-level inference.

6.8.3 Long-term translational priorities

1. Conduct multi-center prospective validation under realistic institutional workflow constraints.

2. Evaluate clinician trust, workload, and explanation usability through structured human-factors studies.

3. Provide governance plans for drift monitoring, audit frequency, and maintenance of the safety case.

4. Progressively align the system with institutional and regulatory requirements for clinical decision support.

6.8.4 Suggested Evaluation Extensions

To maintain the quality of dissertation research, extensions to the evaluation are suggested

1. broader confidence-interval reporting with stratified subgroup analyses

2. paired statistical tests for wrapper-level ablations when paired outputs exist

3. explicit error-taxonomy tracking that distinguishes safety-triggered refusals from genuine low-evidence abstentions

4. reporting summaries that include utility, safety and uncertainty rather than utility only.

6.8.5 Recommended Program of Work

A practical next-stage program can be organized into three work packages:

1. Work Package A (Data and annotation): increase the diversity of the benchmark, improve adjudication practices, and release agreement statistics.

2. Work Package B (Model and retrieval): enhance severity conditioning, ranking and calibrated abstention under prescribed safety margins.

3. Work Package C (Human-in-the-loop expert validation): conduct formal studies with clinicians in the loop to gather more expert opinions on usefulness, trust calibration, explanation quality, evidence relevance, and recovery strategies in case of failure. This work package is designed to complement technical benchmark testing by consulting with clinical experts to determine whether the system's outputs are interpretable, safe enough, evidence-based and useful in realistic decision-making scenarios.

This package plan converts the agenda for future research into manageable packages that may be used to plan grant proposals, conduct lab-based research, or engage in multi-institutional collaborative research.

6.9 Final Conclusion

This dissertation shows that we can make headway towards clinically relevant GI MedVQA via constrained, evidence-connected, auditable system design. The key take-home message is not that generation should replace traditional reliability anchors. Instead, generation is useful only in the context of constrained tasks, evaluation policy, and workflow safety measures.

Pragmatically, the thesis offers a defensible bridge from benchmark-oriented VQA to clinician-sensitive multimodal decision support. It provides incremental empirical benefits, clear limits, and a clear domain for future work that can be done without violating claim discipline.

The last implication for future work is methodological: in high-stakes multimodal AI, reliability is a systems attribute. Architecture, data, prompt and policy constraints, retrieval, escalation all must be assessed together. This dissertation provides an initial roadmap for such an integrated evaluation in the case of GI-endoscopy MedVQA and in UC severity-focused clinical reasoning support.

Appendix A. Reproducibility Artifacts and Internal Repository Sources

Artifact ID	Repository path	Used in	Purpose
A1	Prototyping_reformat/DatasetAnalysis/HyperKvasir/HyperKvasir.md	Chapters 2-3	HyperKvasir local report and result summary used for GI visual-grounding and long-tail imbalance analysis.
A2	Prototyping_reformat/DatasetAnalysis/ImageCLEF_MEDVQA_GI_2023/ImageCLEF_MEDVQA_GI_2023.md	Chapters 2-3	ImageCLEF MEDVQA-GI 2023 local report and validation results used for closed-label GI VQA analysis.
A3	Prototyping_reformat/DatasetAnalysis/LIMUC/LIMUC.md; Prototyping_reformat/DatasetAnalysis/LIMUC/4_reporting/out/	Chapters 3-4	LIMUC local report, split details, severity metrics, and frozen reporting outputs used for UC Mayo severity evaluation.
A4	Prototyping_reformat/DatasetAnalysis/Kvasir_VQA/Kvasir_VQA.md	Chapters 2-3	Kvasir-VQA local report and subset analysis used for colonoscopy VQA question-family and answer-format evaluation.
A5	Prototyping_reformat/DatasetAnalysis/Kvasir_VQA_x1/Kvasir_VQA_x1.md	Chapters 2-3	Kvasir-VQA-x1 local report and generative evaluation diagnostics used for large-scale free-text VQA analysis.
A6	Prototyping_reformat/DatasetAnalysis/Kvasir_SEG/Kvasir_SEG.md	Chapter 3	Supporting localization and morphology context used to motivate grounding-aware evaluation.
A7	Prototyping_reformat/DatasetAnalysis/LIMUC/0_dataset_prep/01_build_metadata_images_and_manifests.ipynb	Chapter 4	LIMUC metadata, image manifest, split construction, and split-freezing notebook.
A8	Prototyping_reformat/DatasetAnalysis/LIMUC/3_vlm_severity/vlm_zero_shot_mayo.ipynb	Chapter 4	Zero-shot Mayo severity baseline notebook used to evaluate prompt-only severity prediction.
A9	Prototyping_reformat/DatasetAnaly-	Chapter 4	LoRA-based controlled generative severity adaptation notebook used for

	sis/LIMUC/3_vlm_severity/vlm_lora_fine-tune_mayo.ipynb		the main Chapter 4 generative lane.
A10	Prototyping_reformat/chapter5_pico_wrapper/results/eval_pass4_latest/pico_eval.json	Chapter 5	PICO extraction evaluation file used for field-level precision, recall, and F1 results.
A11	Prototyping_reformat/chapter5_pico_wrapper/results/eval_pass4_latest/retrieval_eval.json	Chapter 5	Retrieval evaluation file used for Precision@k, Recall@k, Hit Rate@k, and confidence interval reporting.
A12	Prototyping_reformat/chapter5_pico_wrapper/results/eval_pass4_latest/answer_eval.json	Chapter 5	Answer quality and citation-grounding evaluation file used for claim support, citation coverage, refusal rate, and hallucination proxy metrics.
A13	Prototyping_reformat/chapter5_pico_wrapper/results/wrapper_eval_pass4_latest/run_config.json	Chapter 5	Wrapper run configuration file used to document retrieval backend, reranking settings, query count, and severity-context status.
A14	Prototyping_reformat/chapter5_pico_wrapper/results/chapter5_completion_audit_ch5_freeze_20260306/	Chapter 5	Completion audit and freeze-report directory used to verify that the Chapter 5 pipeline artifacts passed the final audit checks.
A15	Prototyping_reformat/chapter5_pico_wrapper/results/kb_build_pass4_latest/kb_manifest.json	Chapter 5	Knowledge-base manifest used to document source files, document count, chunk count, chunking parameters, and retrieval backend availability.
A16	Prototyping_reformat/chapter5_pico_wrapper/results/eval_pass3_ablation/retrieval_ablation_summary.tsv	Chapter 5	Retrieval ablation summary used to compare backend and reranking configurations.
A17	Prototyping_reformat/chapter5_pico_wrapper/pico_wrapper/schemas.py	Chapter 5	Typed schema definitions for PICO frames, severity results, evidence chunks, citations, wrapper inputs, and wrapper outputs.
A18	Prototyping_reformat/chapter5_pico_wrapper/pico_wrapper/retriever.py	Chapter 5	Retriever implementation used for backend-aware and PICO-conditioned evidence retrieval.
A19	Prototyping_reformat/chapter5_pico_wrapper/pico_wrapper/safety.py	Chapter 5	Safety-policy implementation used for refusal, escalation, disclaimer, and

			policy-bounded answer behavior.
A20	Prototyping_reformat/chapter5_pico_wrapper/pico_wrapper/wrapper.py	Chapter 5	Pipeline orchestration module connecting PICO extraction, retrieval, synthesis, safety checks, and wrapper output generation.
A21	Prototyping_reformat/DatasetAnalysis/LIMUC/1_frozen_encoders	Chapter 4	Frozen-encoder baseline notebooks used for ResNet50, ViT, and CLIP linear baseline comparisons.
A22	Prototyping_reformat/DatasetAnalysis/LIMUC/2_supervised_finetuning	Chapter 4	Supervised fine-tuning notebooks used for ResNet50 and ViT/Swin severity baselines.
A23	Prototyping_reformat/DatasetAnalysis/Kvasir_VQA_x1/2_modeling/09_rag_blip2_eval/01_rag_blip2_eval.ipynb	Chapter 4	Retrieval-supported BLIP-2 evaluation notebook used as the design pattern for later evidence-aware extension.
A24	Thesis/markdown/figures/ch4_representations	Chapter 4	Chapter 4 synchronized figure assets and representation files used for metric comparison, radar/profile, confusion matrix, and domain-shift figures.

References

Ben Abacha A, Hasan SA, Datla V, et al. 2019. "VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019." CEUR Workshop Proceedings Vol-2380. https://ceur-ws.org/Vol-2380/paper_78.pdf.

Borgli H, Thambawita V, Smedsrud PH, et al. 2020. "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy." Scientific Data. https://doi.org/10.1038/s41597-020-00622-y.

Dong W, Shen S, Han Y, et al. 2025. "Generative Models in Medical Visual Question Answering: A Survey." Applied Sciences. https://doi.org/10.3390/app15062983.

Fiaidhi J, Mohammed S, Zezos P. 2022. "An xAI Thick Data Assisted Caption Generation for Labeling Severity of Ulcerative Colitis Video Colonoscopy." In 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI), pp. 647-652. IEEE. https://doi.org/10.1109/ICHI54592.2022.00131.

Fiaidhi J, Mohammed S, Zezos P. 2023. "Siamese Neural Network for Labeling Severity of Ulcerative Colitis Video Colonoscopy: A Thick Data Approach." In Intelligent Systems and Applications (IntelliSys 2022), Lecture Notes in Networks and Systems, vol. 542, pp. 124-135. Springer, Cham. https://doi.org/10.1007/978-3-031-16072-1_9

Gautam S, Riegler MA, Halvorsen P. 2025a. "Kvasir-VQA-x1: A Multimodal Dataset for Medical Reasoning and Robust MedVQA in Gastrointestinal Endoscopy." arXiv:2506.09958. https://arxiv.org/abs/2506.09958.

Gautam S, Riegler MOD, Sivertsen KD, Halvorsen P. 2025b. "CLoE: Improving Endoscopic Severity Rating Through Curriculum Learning in Vision Language Models." arXiv:2508.13280. https://arxiv.org/abs/2508.13280.

Gautam S, Storås A, Midoglu C, et al. 2024. "Kvasir-VQA: A Text-Image Pair GI Tract Dataset." arXiv:2409.01437. https://arxiv.org/abs/2409.01437.

Gautam S, Thambawita V, Riegler M, Halvorsen P, Hicks S. 2025c. "Medico 2025: Visual Question Answering for Gastrointestinal Imaging." arXiv:2508.10869. https://arxiv.org/abs/2508.10869.

Hashash JG, Farraye FA, Wang Y, et al. 2024. "Inter- and Intraobserver Variability on Endoscopic Scoring Systems in Crohn's Disease and Ulcerative Colitis: A Systematic Review and Meta-Analysis." Inflammatory Bowel Diseases. https://pubmed.ncbi.nlm.nih.gov/38547325/.

He X, Zhang Y, Mou L, et al. 2020. "PathVQA: 30000+ Questions for Medical Visual Question Answering." arXiv:2003.10286. https://arxiv.org/abs/2003.10286.

Hicks SA, Strumke I, Thambawita V, et al. 2022. "On evaluation metrics for medical applications of artificial intelligence." Scientific Reports. https://www.nature.com/articles/s41598-022-09954-8.

Hu EJ, Shen Y, Wallis P, et al. 2021. "LoRA: Low-Rank Adaptation of Large Language Models." arXiv:2106.09685. https://arxiv.org/abs/2106.09685.

Hu Y, Li T, Lu Q, et al. 2024. "OmniMedVQA: A New Large-Scale Comprehensive Evaluation Benchmark for Medical LVLM." arXiv:2402.09181. https://arxiv.org/abs/2402.09181.

Huang X, Wang N, Liu H, Tang X, Zhou Y. 2025. "MedVLSynther: Synthesizing High-Quality Visual Question Answering from Medical Documents with Generator-Verifier LMMs." arXiv:2510.25867. https://arxiv.org/abs/2510.25867.

ImageCLEF. 2023. "ImageCLEFmed MEDVQA-GI 2023 task page." https://www.imageclef.org/2023/medical/vqa.

ImageCLEF. 2024. "ImageCLEFmed VQA 2024 Task Page." https://www.imageclef.org/2024/medical/vqa.

ImageCLEF. 2025. "ImageCLEFmed MEDVQA 2025 task page." https://www.imageclef.org/2025/medical/vqa.

Jiang S, Wang Y, Song S, et al. 2025. "OmniV-Med: Scaling Medical Vision-Language Model for Universal Visual Understanding." arXiv:2504.14692. https://arxiv.org/abs/2504.14692.

Karim AHMR, Uzuner O. 2025. "MasonNLP at MEDIQA-WV 2025: Multimodal Retrieval-Augmented Generation with Large Language Models for Medical VQA." ClinicalNLP. https://aclanthology.org/2025.clinicalnlp-1.10/.

Lau JYC, Gayen S, Ben Abacha A, et al. 2018. "A dataset of clinically generated visual questions and answers about radiology images." Scientific Data. https://www.nature.com/articles/sdata2018251.

Lee J, Yoon W, Kim S, et al. 2019. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." arXiv:1901.08746. https://arxiv.org/abs/1901.08746.

Li C, Wong C, Zhang S, et al. 2023. "LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day." arXiv:2306.00890. https://arxiv.org/abs/2306.00890.

Li J, Li D, Savarese S, Hoi SCH. 2023. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." arXiv:2301.12597. https://arxiv.org/abs/2301.12597.

Li R, Liu L, Xie Q, et al. 2025. "Towards Medical Visual Question Answering with Large Multimodal Models." arXiv:2501.07109. https://arxiv.org/abs/2501.07109.

Lim DYZ, Basha A, Ku A, et al. 2025. "Vision-language large learning model, GPT4V, outperforms machine learning and deep learning methods in grading bowel preparation quality in outpatient colonoscopies." BMJ Open Gastroenterology 12:e001496. https://pmc.ncbi.nlm.nih.gov/articles/PMC11911458/.

Lin S, Kryściński W, Wu D, et al. 2021. "Medical Visual Question Answering: A Survey." arXiv:2111.10056. https://arxiv.org/abs/2111.10056.

Lin Z, Zhang D, Tao Q, et al. 2023. "Medical visual question answering: A survey." Artificial Intelligence in Medicine. https://doi.org/10.1016/j.artmed.2023.102611.

Liu B, Zhan LM, Xu L, Ma L, Yang Y, Wu XM. 2021. "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. IEEE. https://doi.org/10.1109/ISBI48211.2021.9434010.

Liu H, Li C, Wu Q, Lee YJ. 2023. "Visual Instruction Tuning (LLaVA)." arXiv:2304.08485. https://arxiv.org/abs/2304.08485.

Lu J, Batra D, Parikh D, Lee S. 2019. "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks." arXiv:1908.02265. https://arxiv.org/abs/1908.02265.

Luo R, Sun L, Xia Y, et al. 2022. "BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining." arXiv:2210.10341. https://arxiv.org/abs/2210.10341.

MediaEval. 2025. "Medico 2025 task page: VQA with multimodal explanations for GI imaging." https://multimediaeval.github.io/editions/2025/tasks/medico/.

Mohammed S, Fiaidhi J. 2024. "Generative AI for Evidence-Based Medicine: A PICO GenAI for Synthesizing Clinical Case Reports." In ICC 2024 - IEEE International Conference on Communications, pp. 1503-1508. IEEE. https://doi.org/10.1109/ICC51166.2024.10622271

Moor M, Huang Q, Wu S, et al. 2023. "Med-Flamingo: a Multimodal Medical Few-shot Learner." arXiv:2307.15189. https://arxiv.org/abs/2307.15189.

Murino A, Rimondi A. 2023. "Automated artificial intelligence scoring systems for the endoscopic assessment of ulcerative colitis: How far are we from clinical application?" Gastrointestinal Endoscopy. https://pubmed.ncbi.nlm.nih.gov/36509572/.

Murtaza N, Munsif S, Cuadros M, et al. 2023. "Overview of ImageCLEFmedical 2023 - Medical Visual Question Answering for Gastrointestinal Tract." CEUR-WS Vol-3497. https://ceur-ws.org/Vol-3497/paper-107.pdf.

Nguyen H-D, Dang M-A, Le M-T, Le M-T. 2025. "MedXplain-VQA: Multi-Component Explainable Medical Visual Question Answering." arXiv:2510.22803. https://arxiv.org/abs/2510.22803.

Ozawa T, Ishihara S, Fujishiro M, et al. 2020. "Novel Computer-Aided Diagnosis System for Endoscopic Disease Activity in Patients with Ulcerative Colitis." Gastroenterology 158(8):2150-2157.e3. https://www.gastrojournal.org/article/S0016-5085%2820%2930212-2/fulltext.

Polat G, Kani HT, Ergenc I, et al. 2022. "Labeled Images for Ulcerative Colitis (LIMUC) Dataset." Zenodo. https://zenodo.org/records/5827695. Secondary resource: https://github.com/wanghaining/ulcerative_colitis.

Radford A, Kim JW, Hallacy C, et al. 2021. "Learning Transferable Visual Models From Natural Language Supervision." arXiv:2103.00020. https://arxiv.org/abs/2103.00020.

Rieff M, Varma M, Rabow O, et al. 2025. "SMMILE: An Expert-Driven Benchmark for Multimodal Medical In-Context Learning." arXiv:2506.21355. https://arxiv.org/abs/2506.21355.

Safwan I, Shaikh MA, Haaris M, Khan R, Tahir MA. 2025. "Multi-Task Learning for Visually Grounded Reasoning in Gastrointestinal VQA." arXiv:2511.04384. https://arxiv.org/abs/2511.04384.

Sial M, Fatima M, Nawaz K, et al. 2025. "Path-RAG: Knowledge-Based Explainable Medical VQA with Large Language Models." Proceedings of Machine Learning Research 259. https://proceedings.mlr.press/v259/sial25a.html.

Simula. n.d. "Kvasir-VQA-x1 GitHub repository." https://github.com/simula/Kvasir-VQA-x1.

Simula. 2024. "ImageCLEFmed-MEDVQA-GI-2024 repository." https://github.com/simula/ImageCLEFmed-MEDVQA-GI-2024.

Simula Datasets. n.d. "Kvasir-VQA dataset page." https://datasets.simula.no/kvasir-vqa/.

Smedsrud PH, Thambawita V, Hicks SA, et al. 2021. "Kvasir-Capsule, a video capsule endoscopy dataset." Scientific Data. https://www.nature.com/articles/s41597-021-00920-z.

Stidham RW, Liu W, Bishu S, et al. 2019. "Performance of a Deep Learning Model vs Human Reviewers in Grading Endoscopic Disease Severity of Patients With Ulcerative Colitis." JAMA Network Open 2(5):e193963. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2733432.

Takenaka K, Ohtsuka K, Fujii T, et al. 2023. "Development and Validation of a Deep Neural Network for Accurate Evaluation of Endoscopic Images From Patients With Ulcerative Colitis." Journal of Crohn's and Colitis 17(4):463-472. https://academic.oup.com/ecco-jcc/article/17/4/463/6762568.

Tan H, Bansal M. 2019. "LXMERT: Learning Cross-Modality Encoder Representations from Transformers." arXiv:1908.07490. https://arxiv.org/abs/1908.07490.

Yan Q, He X, Yue X, Wang XE. 2024. "Worse than Random? An Embarrassingly Simple Probing Evaluation of Large Multimodal Models in Medical VQA." arXiv:2405.20421. https://arxiv.org/abs/2405.20421.

Yao H, Tewari AK, Morais M, et al. 2023. "Novel deep learning-based computer-aided diagnosis system for predicting inflammatory activity in ulcerative colitis: a prospective multicentre study." Gastrointestinal Endoscopy 97(2):330-339.e1. https://pubmed.ncbi.nlm.nih.gov/35985375/.

Yim W-w, Ben Abacha A, Yetisgen M, Xia F. 2025. "Overview of the MEDIQA-WV 2025 Shared Task on Woundcare Visual Question Answering." ClinicalNLP. https://aclanthology.org/2025.clinicalnlp-1.3/.

Yip SL, He S, Nie Y, et al. 2025. "MedBookVQA: A Systematic and Comprehensive Medical Benchmark Derived from Open-Access Book." arXiv:2506.00855. https://arxiv.org/abs/2506.00855.

Yu S, Wang H, Wu J, et al. 2025. "MedFrameQA: A Multi-Image Medical VQA Benchmark for Clinical Reasoning." arXiv:2505.16964 (revised 2026). https://arxiv.org/abs/2505.16964.

Zhang X, Wu C, Zhao Z, et al. 2023. "PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering." arXiv:2305.10415. https://arxiv.org/abs/2305.10415.