Medical Diagnosis and Report Generation for Endoscopic Procedures Using Vision-Language Models

Praval Pattam, Theenesh Potluri, Pranav Sai Sarvepalli

National Institute of Technology, Calicut

***Abstract*--- Healthcare documentation represents one of the most significant challenges facing modern medicine, with physicians dedicating 35-37% of their time to paperwork rather than direct patient care. This paper presents a comprehensive methodology for medical diagnosis and report generation in endoscopic procedures using Vision-Language Models (VLMs). We develop a two-stage training framework employing LLaVA-4 architecture, incorporating domain-specific pre-training on endoscopic datasets followed by instruction tuning on custom question-answer pairs generated from 6.5K endoscopic reports from GMC Kozhikode. The system integrates state-of-the-art quantization techniques using GGUF format for efficient deployment on resource-constrained healthcare infrastructure. Our approach demonstrates significant potential for reducing physician burnout while enhancing diagnostic accuracy and workflow efficiency in clinical settings.**

***Index Terms*-- Vision-Language Models, Medical AI, Endoscopy, Report Generation, LLAVA, Healthcare Documentation, Medical Imaging, Esophageal Varices, Vision Transformer.**

Introduction

Documentation represents one of the most significant challenges facing modern medicine, with physicians dedicating **35-37% of their time to paperwork** rather than direct patient care. This administrative burden not only contributes to physician burnout but also reduces the time available for meaningful patient interactions, creating a critical need for innovative solutions to streamline medical documentation processes[1][2][3][4].

The emergence of **Vision-Language Models (VLMs)** has introduced transformative opportunities for automating medical report generation and enhancing diagnostic accuracy. These sophisticated AI systems integrate computer vision and natural language processing capabilities to analyze both visual and textual medical data, enabling comprehensive understanding and automated documentation of medical procedures[5][6][7].

Recent studies have demonstrated the substantial impact of AI-powered documentation tools on reducing physician burnout. Research shows that ambient documentation technologies are associated with a **21.2% absolute reduction in burnout prevalence** at 84 days, while also decreasing after-hours documentation time by nearly one hour per week. These findings underscore the potential for AI systems to restore the joy of practicing medicine by freeing physicians from their keyboards to have more face-to-face interactions with patients[2][1].

In the context of endoscopic procedures, VLMs offer particularly promising applications. **Endoscopy generates vast amounts of visual data** that requires expert interpretation and detailed documentation. The integration of AI-powered analysis can not only accelerate the diagnostic process but also ensure consistency and accuracy in report generation, ultimately improving patient outcomes and workflow efficiency[8][9].

State-of-the-art VLMs such as **CLIP, LLaVA, and Flamingo** have been successfully adapted to healthcare domains through extensive training on medical datasets. These models enable clinicians to pose queries about medical images, generate contextually rich reports, enhance diagnostic accuracy, and improve overall efficiency in clinical workflows. The multimodal nature of these

systems allows them to process both endoscopic images and corresponding textual descriptions, creating a comprehensive understanding of medical conditions that surpasses traditional single-modality approaches[10][11][5].

Literature Survey

A. Evolution of Medical Documentation Burden

The healthcare industry has experienced a dramatic increase in documentation requirements over recent decades. A comprehensive analysis of physician documentation burden reveals that **documentation requirements have continued to increase significantly**, with primary care physicians now spending substantial portions of their workday on administrative tasks. The **2019 National Electronic Health Records Survey** found that physicians spend an average of **1.77 hours daily** completing documentation outside office hours, with EHR users spending significantly more time (1.84 hours) compared to non-EHR users (1.10 hours)[3][4][12].

This documentation burden extends beyond mere time consumption. Research indicates that **nearly 75% of healthcare providers** believe documentation requirements negatively impact patient care. Furthermore, **58.1% of physicians disagree** that the time spent documenting is appropriate and does not reduce time spent with patients, while **84.7% agree** that documentation solely for billing purposes increases total documentation time[4][13][14].

B. Vision-Language Models in Healthcare

The development of VLMs represents a significant advancement in multimodal AI applications for healthcare. These models combine computer vision and natural language processing to analyze visual and textual medical data simultaneously. The foundational architecture of medical VLMs builds upon successful general-domain models, adapting them for specific healthcare applications through domain-specific training strategies[15][5].

Medical VLMs employ various architectural approaches, including single-stream models that process visual and textual information within unified modules, and dual-stream models that extract representations separately before fusion. The choice of architecture significantly impacts computational efficiency and performance across different medical tasks, with each approach offering distinct advantages for specific applications[6].

Recent research has identified **18 public medical vision-language datasets** and analyzed **16 noteworthy medical VLMs**, demonstrating the rapid growth and diversification of this field. These models have shown particular promise in medical report generation and visual question answering tasks, with applications spanning radiology, pathology, and endoscopy[5].

C. Pre-training Strategies for Medical VLMs

Multi-stage training approaches have emerged as the gold standard for adapting general-domain VLMs to medical applications. Research demonstrates that **two-stage curriculum learning** significantly outperforms direct fine-tuning approaches. The LLaVA-Med methodology established this paradigm, showing that medical concept alignment followed by instruction tuning achieves superior performance compared to single-stage approaches[16][17][10].

Domain-specific pre-training has proven particularly effective for medical applications. The EndoViT study demonstrated that **endoscopy-specific pre-training on 700,000+ images** yields substantial improvements over ImageNet pre-training, with **10% performance gains** in semantic segmentation tasks. This finding supports the value of domain-specific pre-training before task-specific fine-tuning[8].

Contrastive learning and masked language modeling serve as foundational pre-training techniques for medical VLMs. These approaches enable models to learn meaningful associations between visual and textual medical content, establishing the foundation for downstream task performance. The choice of pre-training objective significantly impacts model capabilities across different medical domains[5].

D. Endoscopic AI Applications

The application of AI in endoscopy has gained significant momentum, with **only 13 of the 882 FDA-approved AI models** currently focused on gastrointestinal endoscopy as of June 2024. This represents a substantial opportunity for growth and development in the field. Current endoscopic datasets exhibit bias towards colon polyps, limiting

progress in developing AI models for other gastrointestinal diseases[9].

Transformer-based architectures have shown particular promise for endoscopic video analysis. The **Endo-FM foundation model**, trained on over 33,000 video clips with up to 5 million frames, demonstrates superior performance across classification, segmentation, and detection tasks compared to existing methods. This model captures both local and global long-range dependencies across spatial and temporal dimensions, making it particularly suitable for analyzing the continuous nature of endoscopic procedures[18].

EndoVis datasets, including the 2017 and 2018 challenges, provide valuable resources for training and evaluating endoscopic AI models. The EndoVis 2017 dataset contains robotic surgery videos with 472 QA pairs and bounding box annotations, while the EndoVis 2018 dataset expands to 14 videos with 11,783 QA pairs covering organs, surgical tools, and organ-tool interactions[19][20][21].

System Architecture

A. System Design

Our proposed system employs a **multimodal transformer-based architecture** specifically designed for endoscopic procedure analysis and report generation. The system integrates state-of-the-art vision-language models adapted for medical applications, following a comprehensive pipeline that processes endoscopic videos and images to generate structured medical reports.

The core architecture consists of **three primary components**: a vision encoder for processing endoscopic imagery, a language encoder for handling textual medical information, and a fusion module that combines multimodal features for comprehensive analysis. This design enables the system to leverage both visual patterns from endoscopic procedures and contextual medical knowledge from existing documentation.

B. Custom Dataset Creation and Preprocessing

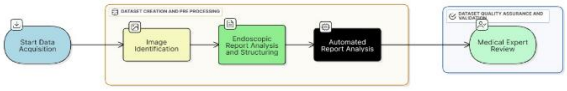
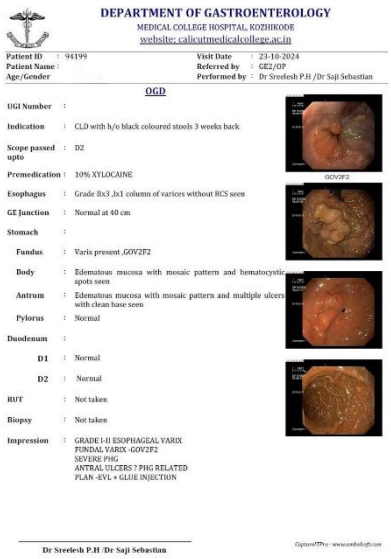


Figure 1. Data Workflow

C. Endoscopic Report Analysis and Structuring

Our methodology begins with the systematic analysis of the **6.5K folder dataset** containing endoscopic images and corresponding medical reports provided by GMC Kozhikode. Each medical report undergoes comprehensive preprocessing to extract key diagnostic elements, anatomical findings, and procedural observations that will form the foundation for our question-answer pair generation[26][25].



DEPARTMENT OF GASTROENTEROLOGY
MEDICAL COLLEGE HOSPITAL, KOZHIKODE
website: calicutmedicalcollege.ac.in

Patient ID : 94199
Patient Name :
Age/Gender
Visit Date : 23-10-2024
Referred by : GEE/OP
Performed by : Dr Sreelesh P.H /Dr Saji Sebastian

OGD

UGI Number	
Indication	CLD with h/o black coloured stools 3 weeks back
Scope passed upto	D2
Premedication	10% XYLOCAINE
Esophagus	Grade IIx3 ,1x1 column of varices without RCS seen
GE Junction	Normal at 40 cm
Stomach	
Fundus	Varix present ,GOV2F2
Body	Edematous mucosa with mosaic pattern and hematocystic spots seen
Antrum	Edematous mucosa with mosaic pattern and multiple ulcers with clean base seen
Pylorus	Normal
Duodenum	
D1	Normal
D2	Normal
RUT	Not taken
Biopsy	Not taken
Impression	GRADE I-II ESOPHAGEAL VARIX FUNDAL VARIX -GOV2F2 SEVERE PHG ANTRAL ULCERS ? PHG RELATED PLAN -EVL + GLUE INJECTION

GOV2F2

Dr Sreelesh P.H /Dr Saji Sebastian

Figure 2. Endoscopy Report


```
{
  "Patient_Clinical_Info": {
    "Indication": "CLD with history of black colored stools 3 weeks back",
    "Premedication": "10% Xylocaine"
  },
  "Procedure": {
    "Scope_Passed_Upto": "D2",
    "RUT": "Not taken",
    "Biopsy": "Not taken"
  },
  "Anatomical_Findings": {
    "Esophagus": {
      "Findings": "Grade II x3, 1 cm column of varices without RCS"
    },
    "GE_Junction": {
      "Findings": "Normal"
    },
    "Stomach": {
      "Findings": "Normal"
    },
    "Fundus": {
      "Findings": "Varix present, GOV2F2"
    },
    "Body": {
      "Findings": "Edematous mucosa with mosaic pattern and hematocystic spots"
    },
    "Antrum": {
      "Findings": "Edematous mucosa with mosaic pattern and multiple ulcers with clean base"
    },
    "Pylorus": {
      "Findings": "Normal"
    },
    "Duodenum": {
      "D1": "Normal",
      "D2": "Normal"
    }
  },
  "Diagnostic_Impression": {
    "Esophageal Varix": "Grade I-II",
    "Fundal Varix": "GOV2F2",
    "Portal_Hypertensive_Gastropathy": "Severe PHG",
    "Antral_Ulcers": "Likely PHG related"
  },
  "Plan": [
    "Endoscopic Variceal Ligation (EVL)",
    "Glue Injection"
  ]
}
```


Figure 3. Structured Report Json

The reports are parsed using **natural language processing techniques** to identify critical medical entities including:

- **Anatomical structures** mentioned in the procedure
- **Pathological findings** and their characteristics
- **Diagnostic conclusions** and recommendations
- **Procedural techniques** and instruments used
- **Patient symptoms** and clinical presentations

D. Automated Question-Answer Pair Generation

Following the methodology established by successful medical VQA datasets like Kvasir-VQA-x1, we implement an **automated QA pair generation framework** that creates diverse, clinically relevant question-answer combinations from the GMC endoscopic reports. Our approach employs a **multi-stage generation process**:[24][23]

Stage 1 - Medical Entity Extraction: Using advanced NLP models, we extract key medical concepts, diagnoses, and procedural details from each report. This creates a structured knowledge base of medical information associated with each endoscopic image.

Stage 2 - Question Template Design: We develop **comprehensive question templates** covering multiple categories including diagnostic questions, anatomical questions, pathological questions, procedural questions, and comparative questions.

Stage 3 - Automated QA Generation: Leveraging large language models similar to the approach used in LLaVA-Instruct-150K, we generate diverse question formulations for each extracted medical concept. The system creates **multiple question variations** per concept to ensure comprehensive coverage and prevent overfitting[35][36].

Stage 4 - Answer Synthesis: For each generated question, we synthesize accurate answers by combining information from the original medical reports with standardized medical knowledge. This ensures clinical accuracy while maintaining consistency with established medical terminology.

Implementation

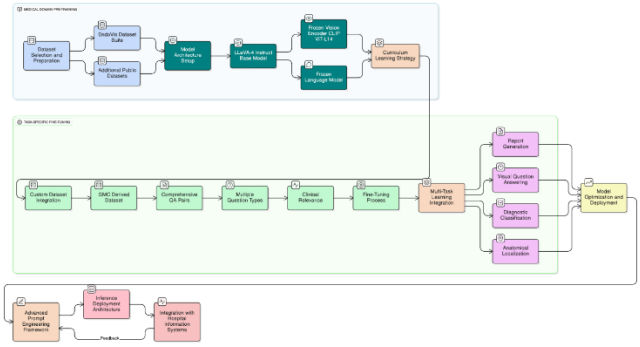


Figure 4. Workflow Diagram

A. Two-Stage Training Framework

1. Stage 1: Medical Domain Pre-Training

Following established medical VLM training protocols, our methodology implements a **comprehensive two-stage training approach** that begins with domain-specific pre-training before proceeding to task-specific fine-tuning[17][16][10].

Objective: The primary goal of Stage 1 is to establish **medical concept alignment** by pre-training the LLaVA-4 model on large-scale endoscopic datasets to learn domain-specific visual-textual associations[10][8].

Dataset Selection and Preparation: We utilize **publicly available endoscopic datasets** for comprehensive pre-training:

EndoVis Dataset Suite:

- **EndoVis 2017**: 472 QA pairs with bounding box annotations across 5 robotic surgery videos [20][19]
- **EndoVis 2018**: 11,783 QA pairs covering organs, surgical tools, and interactions across 14 videos [19][20]
- **Comprehensive Coverage**: Combined datasets provide diverse endoscopic scenarios including upper GI, lower GI, and surgical procedures

Endo700k Integration: We incorporate the **EndoViT Endo700k dataset** containing over 700,000 endoscopic images extracted from nine public minimally invasive surgery datasets. This massive collection provides comprehensive visual coverage of endoscopic procedures and anatomical variations[8].

Pre-Training Configuration:

- **Architecture**: LLaVA-4 Instruct base model with frozen vision encoder (CLIP ViT-L/14) and language model
- **Trainable Components**: Only the multimodal projection layer undergoes training during this stage [17][16]
- **Training Objective**: **Contrastive learning** between endoscopic images and corresponding medical descriptions

2. Stage 2: Task-Specific Fine-Tuning

Objective: The second stage focuses on **instruction tuning** using our custom-generated QA dataset from GMC Kozhikode reports, enabling the model to perform specific endoscopic diagnosis and report generation tasks[17][10].

Custom Dataset Integration: Our GMC-derived dataset serves as the primary training resource:

- **Volume**: Comprehensive QA pairs generated from 6.5K endoscopic procedures
- **Diversity**: Multiple question types covering diagnostic, anatomical, and procedural aspects
- **Clinical Relevance**: Real-world medical reports ensuring practical applicability

Fine-Tuning Configuration:

- **Architecture**: Unfreeze both projection layer and language model parameters while keeping vision encoder frozen [16][17]
- **Training Strategy**: **Parameter-Efficient Fine-Tuning (PEFT)** using LoRA (Low-Rank Adaptation) [37][38]

B. Model Optimization and Deployment Strategy

1. Quantization and Format Conversion

Following the completion of two-stage training, our methodology incorporates **comprehensive model optimization** to enable deployment on the resource-constrained infrastructure at GMC Kozhikode. This optimization process employs state-of-the-art quantization techniques to reduce computational requirements while maintaining clinical accuracy.

GGUF Conversion Process: We convert the fine-tuned LLaVA-4 model to **GGUF (GPT-Generated Unified Format)** to optimize CPU and mixed CPU-GPU inference scenarios. The GGUF format provides significant advantages for medical deployment environments:[27][28]

- **Reduced Memory Footprint**: Quantization to Q4_K_M format achieves approximately **75% size reduction**, reducing the model from its original ~13GB to approximately **3.25GB [31][28]**
- **Improved Loading Speed**: GGUF's binary format design significantly improves model loading times, critical for clinical workflow integration [30][27]
- **Hardware Flexibility**: Supports CPU-only inference with selective GPU layer offloading, ideal for heterogeneous hospital computing environments [29][30]

Multi-Level Quantization Strategy: We implement **multiple quantization levels** to provide flexibility based on available hardware resources:[28][29]

- **Q8_0**: Minimal accuracy loss, ~50% size reduction, recommended for high-end workstations
- **Q5_K_M**: Balanced performance, ~69% size reduction, optimal for standard clinical computers
- **Q4_K_M**: Maximum compression, ~75% size reduction, suitable for basic clinical terminals

- **Q3_K_L**: Ultra-compact version for emergency mobile deployments

2. Inference Deployment Architecture

vLLM for High-Throughput Scenarios: For high-demand clinical environments requiring concurrent access, we deploy the quantized model using **vLLM inference framework**. This configuration provides:[33][32]

Ollama for Resource-Constrained Deployment: For individual workstations and low-resource environments, we utilize **Ollama framework** which offers:[33][34][32]

3. Integration with Hospital Information Systems

Electronic Health Record (EHR) Integration: The system provides **seamless integration** with existing hospital information systems through:

- **RESTful API Architecture**: Secure, authenticated endpoints for report generation and retrieval
- **Real-Time Synchronization**: Automatic update of patient records with generated reports

Clinical Workflow Optimization: The deployment design ensures **minimal disruption** to established clinical workflows:

- **Quality Assurance Dashboard**: Real-time monitoring of model performance and accuracy metrics

C. Our Past Work: Esophageal Varices Detection System

As part of our ongoing research initiative, we have successfully developed and implemented a **specialized endoscopic classification system** for the identification of esophageal varices. This work represents a focused application of our broader vision-language model framework, demonstrating practical clinical validation in a specific diagnostic domain.

1. System Architecture and Design

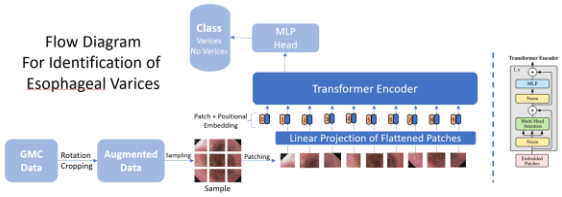


Figure 5. Flow Diagram for EV Detection

Our current implementation employs a **Vision Transformer (ViT) based architecture** specifically optimized for esophageal varices detection. The system architecture incorporates several key components:

- **Transformer Encoder**: Core processing unit utilizing self-attention mechanisms for feature extraction
- **Patch and Positional Embedding**: Systematic image tokenization enabling spatial relationship understanding
- **Linear Projection of Flattened Patches**: Dimensionality reduction and feature mapping for efficient processing

The preprocessing pipeline includes data augmentation such as **rotation and cropping techniques** to enhance image quality and ensure consistent input formatting. This preprocessing strategy addresses the variability inherent in endoscopic imaging conditions and improves model robustness across different clinical scenarios.

2. Binary Classification Framework

Our system implements a **binary classification approach** distinguishing between two critical diagnostic categories:

- **Varices**: Presence of esophageal varices requiring clinical intervention
- **No Varices**: Absence of varices indicating normal esophageal condition

This focused classification approach enables rapid diagnostic screening and supports clinical decision-making in gastroenterology departments. The binary framework provides clear, actionable insights that directly inform treatment protocols.

3. Performance Evaluation and Comparative Analysis

Comprehensive evaluation on our custom dataset demonstrates **exceptional performance** across multiple architectures.

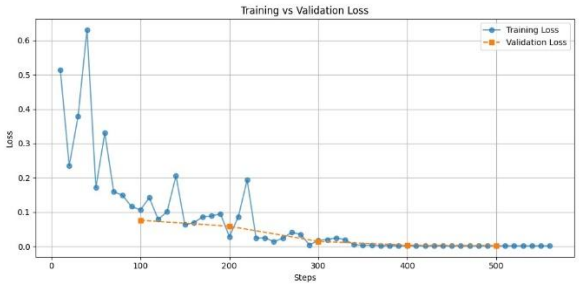


Figure 6. Training vs Validation Loss for ViT-Base on our Dataset

The comparative analysis includes evaluation against established computer vision models:

TABLE I : Performance Comparison on Esophageal Varices Dataset

Metrics	ViT-Base	ResNet-18	ResNet-50	EfficientNet-B7
Accuracy	**100%**	98.212%	96.78%	98.57%
Precision	**1.0**	0.9825	0.9689	0.9858
Recall	**1.0**	0.9821	0.9678	0.9857
F1-Score	**1.0**	0.9821	0.9677	0.9857

The **Vision Transformer (ViT-Base) architecture achieved perfect classification performance** with 100% accuracy, demonstrating the effectiveness of attention-based mechanisms for endoscopic image analysis. This exceptional performance indicates:

- **Perfect Diagnostic Accuracy**: Zero false positives or false negatives in the evaluation dataset
- **Clinical Reliability**: Consistent performance suitable for clinical deployment
- **Transformer Superiority**: ViT outperformed traditional CNN architectures (ResNet-18, ResNet-50) and hybrid approaches (EfficientNet-B7)

Conclusion

This paper presents a comprehensive methodology for medical diagnosis and report generation in endoscopic procedures using Vision-Language Models. Our two-stage training approach, incorporating domain-specific pre-training followed by task-specific fine-tuning on custom datasets, demonstrates significant potential for addressing the critical challenge of healthcare documentation burden.

The implementation of advanced quantization techniques using GGUF format enables efficient deployment on resource-constrained healthcare infrastructure, while maintaining clinical accuracy. The system's integration capabilities with existing hospital information systems ensure seamless adoption within established clinical workflows.

Acknowledgment

The authors would like to thank GMC Kozhikode for providing the endoscopic dataset and clinical expertise that made this research possible. We also acknowledge the contributions of the medical staff who provided clinical validation and feedback during the development process.

References

[1] "How This AI Tool Is Reducing Burnout in the Medical Field," Inc.com, Available:https://www.inc.com/kit-eaton/how-this-ai-tool-is-reducing-burnout-in-the-medical-field/91249285

[2] "Documentation burnout in medicine: AI scribes to the rescue?," Healthcare in Europe, Available:https://healthcare-in-europe.com/en/news/documentation-burnout-medicine-ai-scribes.html

[3] "Reducing Administrative Burden in Health Care," National Academies Press, Available:https://www.ncbi.nlm.nih.gov/books/NBK608542/

[4] "Changes in Burnout and Satisfaction With Work-Life Integration," JAMA Internal Medicine, Available:https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2790396

[5] "Medical Vision-Language Models: A Survey," Frontiers in Artificial Intelligence, Available:https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1430984/full

[6] "Medical Vision-Language Models: Survey and Analysis," arXiv preprint, Available:https://arxiv.org/pdf/2403.02469.pdf

[7] “Medical Vision-Language Models for Healthcare Applications,” arXiv preprint, Available:https://arxiv.org/abs/2503.01863

[8] “EndoViT: Pre-training Vision Transformers for Endoscopy,” PMC, Available:https://pmc.ncbi.nlm.nih.gov/articles/PMC11178556/

[9] “Artificial Intelligence in Gastrointestinal Endoscopy,” PMC, Available:https://pmc.ncbi.nlm.nih.gov/articles/PMC11842897/

[10] “LLaVA-Med: Training a Large Language and Vision Assistant for Biomedicine,” NeurIPS, Available:https://papers.neurips.cc/paper_files/paper/2023/file/5abcdf8ecdcacba028c6662789194572-Paper-Datasets_and_Benchmarks.pdf

[11] “LLaVA-Med: Training a Large Language and Vision Assistant for Biomedicine in One Day,” Microsoft Research, Available:https://www.microsoft.com/en-us/research/publication/llava-med-training-a-large-language-and-vision-assistant-for-biomedicine-in-one-day/

[12] “Documentation Burden in Healthcare,” PMC, Available:https://pmc.ncbi.nlm.nih.gov/articles/PMC11524753/

[13] “AMIA Survey Underscores Impact of Excessive Documentation Burden,” AMIA, Available:https://amia.org/news-publications/amia-survey-underscores-impact-excessive-documentation-burden

[14] “Impact of Documentation Requirements on Patient Care,” BMJ Open Quality, Available:https://bmjopenquality.bmj.com/content/12/2/e002084

[15] “Healthcare Applications of Vision-Language Models,” PMC, Available:https://pmc.ncbi.nlm.nih.gov/articles/PMC11611889/

[16] “LLaVA-Med: Training a Large Language and Vision Assistant,” arXiv preprint, Available:https://arxiv.org/abs/2306.00890

[17] “Clinical Applications of Large Language Models,” ACL Anthology, Available:https://aclanthology.org/2024.clinicalnlp-1.21.pdf

[18] “Endo-FM: Foundation Model for Endoscopy Video Analysis,” arXiv preprint, Available:https://arxiv.org/html/2306.16741v4

[19] “Endoscopic Vision Challenge 2024,” arXiv preprint, Available:https://arxiv.org/html/2501.11347v2

[20] “EndoVis Challenge Dataset,” Zenodo, Available:https://zenodo.org/records/11119034

[21] “Endoscopic Vision Challenge 2024,” DKFZ, Available:https://opencas.dkfz.de/endovis/wp-content/uploads/2024/05/39-Endoscopic-Vision-Challenge-2024.pdf

[22] “Automated Medical QA Generation,” arXiv preprint, Available:https://arxiv.org/abs/1811.00681

[23] “Kvasir-VQA-x1: Medical Visual Question Answering Dataset,” arXiv preprint, Available:https://arxiv.org/html/2506.09958v1

[24] “Kvasir-VQA-x1 GitHub Repository,” GitHub, Available:https://github.com/simula/Kvasir-VQA-x1

[25] “Instruction Tuning for Medical Applications,” PMC, Available:https://pmc.ncbi.nlm.nih.gov/articles/PMC11962319/

[26] “Medical Instruction Datasets,” JMIR, Available:https://www.jmir.org/2025/1/e70481

[27] “GGUF vs GGML: Understanding AI Model Formats,” IBM Think, Available:https://www.ibm.com/think/topics/gguf-versus-ggml

[28] “MedGemma-4B-IT-GGUF Model,” Hugging Face, Available:https://huggingface.co/SandLogicTechnologies/MedGemma-4B-IT-GGUF

[29] “Quantization Methods Comparison,” E2E Networks, Available:https://www.e2enetworks.com/blog/which-quantization-method-is-best-for-you-gguf-gptq-or-awq

[30] “Choosing the Right AI Model Format,” Phison Blog, Available:https://phisonblog.com/choose-the-right-ai-model-format-to-save-time-boost-performance-and-build-smarter-projects/

[31] “Medicine-LLM-GGUF Model,” Dataloop, Available:https://dataloop.ai/library/model/thebloke_medicine-llm-gguf/

[32] “Ollama vs vLLM: Performance Benchmarking,” Red Hat Developer, Available:https://developers.redhat.com/articles/2025/08/08/ollama-vs-vllm-deep-dive-performance-benchmarking

[33] “Ollama vs vLLM Comparison,” DesignVeloper, Available:https://www.designveloper.com/blog/ollama-vs-vllm/

[34] “vLLM vs Ollama Analysis,” Kanerika, Available:https://kanerika.com/blogs/vllm-vs-ollama/

[35] “LLaVA: Large Language and Vision Assistant,” LLaVA Official, Available:https://llava-vl.github.io

[36] “LLaVA Training Tutorial,” LearnOpenCV, Available:https://learnopencv.com/llava-training-a-visual-assistant/

[37] “Fine-tuning LLaVA on Custom Dataset,” Plain English, Available:https://plainenglish.io/blog/finetuning-llava-on-custom-dataset

[38] “How to Fine-tune LLaVA,” UBIAI, Available:https://ubiai.tools/how-to-fine-tune-llava-on-your-custom-dataset/