
GSM-PathEval: A Global South Robustness Benchmark for Telepathology Vision-Language Models

Anshuman Awasthi
Vastram.ai

With
Apart Research

Abstract

Frontier multimodal medical AI systems are primarily evaluated on Western benchmarks featuring pristine, high-resolution pathology scans. However, clinical deployment in the Global South relies on ad-hoc "telepathology", capturing microscope views via budget smartphones and transmitting them over heavily compressed networks like WhatsApp. We introduce ***GSM-PathEval***, a targeted benchmark that programmatically applies real-world infrastructural degradations (eyepiece glare, lens blur, and heavy JPEG quantization) to the gold-standard PathVQA dataset. Using Adaption Labs, we demonstrate a critical "Reality Gap": frontier Vision-Language Models optimized for clean imagery suffer drastic diagnostic accuracy drops $\Delta(A)$ under these simulated field conditions. This performance cliff poses a severe misdiagnosis risk, proving that standard capabilities benchmarks are insufficient for safe medical AI deployment in low-resource environments.

1. Introduction

Problem Statement & Practical Value

Current frontier Vision-Language Models (VLMs) used for medical diagnostics are evaluated almost exclusively on immaculate, high-resolution digital benchmarks. These datasets assume clinical environments equipped with \$100k digital whole-slide scanners, high-bandwidth networks, and perfect lighting.

In the Global South, specifically in rural clinics across India and the broader Asia region, this infrastructure does not exist. Instead, healthcare practitioners rely on "telepathology": capturing images of a microscope eyepiece using low-cost smartphones and transmitting them via heavily compressed mobile chat applications (e.g., WhatsApp) to distant specialists or AI diagnostic assistants.

When unadapted medical VLMs are deployed in these environments, they encounter a severe distribution shift characterized by aggressive compression artifacts, lens blur, and optical glare. Evaluating models solely on pristine western data creates a dangerous "Reality Gap." This project is practically valuable because it builds a testing pipeline modeled on these exact infrastructural constraints, allowing safety teams to audit models for misdiagnosis risks before they are deployed in under-resourced regions.

Threat Model & Failure Modes

The primary failure mode addressed is **feature-erasure induced hallucination**. Cellular pathology relies heavily on micro-features: nuclear atypia, chromatin distribution, and clean membrane boundaries. Structural and transmission degradations (like aggressive JPEG quantization) wipe out these fine-grained high-frequency details. Confronted with ambiguous, blocky artifacts, a fragile VLM may fallback on training priors or text-based prompts, resulting in high-confidence misclassifications (e.g., marking a malignant carcinoma as a benign cyst). This triggers a cascade safety risk, as overworked rural doctors are highly susceptible to automation bias and may trust an invalid AI confirmation.

Main Contributions

Our main contributions are:

1. **The GSM-PathEval Pipeline:** Developed a repeatable, programmatic image-degradation framework that maps specific socioeconomic and infrastructural constraints (cheap mobile sensors, low-bandwidth transmission) onto medical vision datasets.
2. **Quantification of the Reality Gap:** Programmatically extracted a deterministic evaluation subset from the gold-standard PathVQA dataset and executed parallel safety audits using the Adaption Labs platform to expose the operational performance cliff of SOTA vision models.
3. **Open-Source Contextual Auditing Tools:** Released the full self-contained dataset generation script, moving medical AI safety evaluations away from localized baseline capabilities toward context-aware global deployment realities.

2. Related Work

Existing medical VLM evaluations primarily utilize standard benchmarks such as **PathVQA** or **VQA-Med** to measure peak model capability on clean data distributions. While safety-focused benchmarks like *MM-SafetyBench* or *AdvGLUE* do test for adversarial vulnerabilities, they focus on intentional geometric perturbations, typographic attacks, or synthetic noise injections (e.g., salt-and-pepper noise).

The Knowledge Gap

Current safety suites fail to replicate the real-world, systematic, and non-adversarial degradations inherent to Global South healthcare delivery. No existing public benchmark specifically models the downstream effects of consumer-grade mobile optics combined with instant messaging compression matrices on diagnostic tasks.

Why Use GSM-PathEval?

A safety engineer or regional healthcare regulator would utilize our framework over standard SOTA evaluations to discover how an AI system handles real-world operational friction. GSM-PathEval provides vital structural insights that raw capability scores hide, explicitly mapping out how severe network and optical limitations alter the model's clinical risk profile.

3. Methods

Replicability Workflow

To ensure complete reproducibility without local data storage dependencies, our method interfaces directly with the Hugging Face [datasets](#) API to process files via an inline image-transformation stream.

Dataset & Model Infrastructure

We targeted the standard [flaviagiammarino/path-vqa](#) repository on Hugging Face. To establish a deterministic, automated scoring metric suited for a rapid hackathon environment, we programmatically filtered the dataset to select a 50-image subset featuring strict binary (Yes/No) clinical questions (e.g., assessing the definitive presence of malignancy, necrosis, or inflammatory infiltrates).

Key Parameter & Design Decisions

The image degradation pipeline uses the Python **Pillow** library to apply four sequentially stacked mathematical transformations, each justified by empirical field constraints:

- **Brightness Multiplier (*1.3\$)**: Mimics the non-uniform overexposure and blinding ambient reflection common when aligning a smartphone camera manually with an analog microscope eyepiece.
- **Contrast Reduction (*0.8\$)**: Replicates the low dynamic range, color-distortion, and poor sensor quality of budget 5MP mobile camera modules.
- **Gaussian Blur (sigma = 1.2)**: Simulates micro-vibrations, unsteady hand position, lack of fine optical autofocus, and dirty physical lenses.
- **Aggressive Quantization (JPEG Quality = 15)**: Replicates the lossy, high-frequency data destruction forced by low-bandwidth cellular networks (3G/Edge) and the aggressive image resizing pipelines of consumer messaging applications like WhatsApp.

4. Results

Main Findings

Our experiments reveal a stark performance cliff when models shift from idealized environments to simulated field conditions. To evaluate this systematically, we calculate the Drop-in-Accuracy (Delta) using the following formulation:

$$\text{delta}(A) = A(\text{pristine}) - A(\text{rural})$$

Where A(pristine) represents the percentage accuracy on the clean control group, and A (rural) represents the accuracy on the degraded pipeline.

Performance Under Structural Degradation

Evaluation Suite (n=50)	Accuracy Score	Observed Clinical Failure Mode
PathVQA - Pristine (Control)	84%	True positive identification of cellular pathology structure.
GSM-PathEval (Rural Split)	52%	High-confidence false negatives due to blurred cell boundaries.

Performance Drop (Delta (A))	32%	Systematic failure on high-frequency visual diagnostic details.
---	-----	---

Visual Evidence and Interpretations

An analysis of individual evaluation samples from our Adaption Labs output highlights the exact nature of this risk:

As seen during the evaluation runs, when lossy compression alters the continuity of the tissue boundary, the VLM frequently flips its classification from a true positive diagnosis to a definitive negative response, ignoring visible macro-pathology due to the absence of clean micro-margins. The differences are highly robust: the degradation pipeline consistently forces misclassifications across diverse cellular targets.

5. Discussion and Limitations

Implications for AI Safety

The results demonstrate that AI capability metrics published by frontier labs provide a false sense of security regarding medical safety deployment. "Safety" cannot be evaluated independent of infrastructure. When models exhibit high fragility under common, non-adversarial environmental factors like data compression, deploying them without strict input-validation guardrails directly endangers patient lives.

Methodological Limitations

- **Sample Scale Constraint:** Due to the 48-hour time bound of the hackathon, our benchmark cohort was scaled to 50 targeted samples to guarantee clean processing and evaluation runtime.
- **Static Degradation Profiles:** The script applies uniform mathematical transformations. Real-world telepathology involves a wider, dynamic distribution of phone models, motion blurs, and varying stains.
- **Lack of Linguistic Context:** This implementation focuses entirely on visual technical safety. It does not integrate localized linguistic variations (such as medical queries mixed with Hindi or Hinglish idioms) which occur frequently in rural Indian clinical workflows.

Future Work

The natural next step is to expand GSM-PathEval into a multi-tiered safety suite:

1. **Linguistic Perturbation Tier:** Incorporating automated translation layers to evaluate model robustness when facing noisy multilingual prompts.
2. **Sycophancy (Yes-Man) Auditing:** Altering prompt structures to include leading diagnostic biases from clinicians (e.g., *"This looks like high-grade carcinoma, right?"*) to test if the degraded image makes the model more likely to submit to human cognitive errors.

6. Conclusion

GSM-PathEval exposes a critical, unaddressed "Reality Gap" in current multimodal medical AI evaluation. By building an automated degradation framework that maps consumer-grade optics and low-bandwidth network compression onto the PathVQA dataset, we have demonstrated that frontier Vision-Language Models degrade rapidly when evaluated under the real-world operational constraints of the Global South.

Our findings indicate that standard capabilities evaluations are fundamentally insufficient metrics for regional AI safety. If healthcare systems in under-resourced areas deploy medical models without auditing them against context-aware infrastructural degradations, it will inevitably introduce severe automated misdiagnosis risks. Regulators and developers must mandate robust, localized benchmark testing as a non-negotiable step before field deployment.

Code and Data

Code Repository: (<https://github.com/AnshumanAI/gsm-images>)

Data/Datasets: Built programmatically from **flaviagiammarino/path-vqa** on Hugging Face.

References

Flavia Giammarino (2023). *PathVQA: A Dataset for Medical Visual Question Answering in Pathology*. Hugging Face Dataset Repository. Available at:

(<https://huggingface.co/datasets/flaviagiammarino/path-vqa>)

Apart Research (2026). *Global South AIs Hackathon — Asia Track: AI Misdiagnosis Frameworks*. Available at:

(<https://apartresearch.com/sprints/global-south-ais-hackathon>).

Appendix

The following prompt template was passed to the Vision-Language Model alongside both the pristine and degraded images inside the Adaption Labs pipeline to ensure deterministic outputs: You are a professional medical AI assistant specializing in digital pathology.

Examine the provided tissue image carefully.

Answer the following clinical question with exactly one word: 'Yes' or 'No'. Do not add any explanation or preamble.

Question: [Insert Question from PathVQA Dataset Split]

Answer:

LLM Usage Statement

We utilized Gemini to brainstorm appropriate regional safety failure modes, convert clinical constraints into definitive PIL python parameters, and format our text inputs directly into the requested submission layout. All data processing scripts, experimental setups, and evaluation results were independently executed and verified.