SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

Benchmarking Foundation Vision-Language Models for Gastrointestinal Contrast Extravasation Detection on CT

Michael Fei, Creighton University School of Medicine; Christopher Bridge, DPhil; Dania Daye, MD, PhD

Introduction/Background

Vision-Language Models (VLM) are a growing area of AI research, especially in the medical field, where their ability to process images and text aligns naturally with fields such as radiology. Text-only foundation language models have shown surprisingly strong performance on medical tasks, despite not being trained exclusively on medical data. Our study aims to investigate current VLMs in detecting GI extravasation injury, a binary classification task that is conceptually simple but requires relatively complex visuospatial reasoning.

Methods/Intervention

Abdominal CT examinations from Massachusetts General Brigham were used to create a dataset of 500 axial slices with confirmed GI contrast extravasation and 500 anatomically matched control slices. Models received a standardized prompt asking whether GI bleeding was present and responded only "Yes" or "No." Continuous prediction scores were derived from first-token logit probabilities. Evaluated models included MedGemma 1.5 4B, LLaVA-Med v1.5 7B, Llama-3.2-Vision 90B, Gemma-3 27B, Qwen2.5-VL 72B, and InternVL3 78B. Merlin was separately evaluated on 100 positive and 100 control CT volumes. The primary outcome was ROC AUC, with sensitivity, specificity, PPV, F1 score, and accuracy reported at the Youden J threshold.

Results/Outcome

MedGemma 1.5 4B demonstrated the strongest overall performance, achieving an AUC of 0.592 and an accuracy of 0.578. Llama-3.2-Vision 90B performed similarly, with an AUC of 0.588 and an accuracy of 0.572 (Table 1). Merlin achieved an AUC of 0.587, a sensitivity of 0.400, and a specificity of 0.780. The remaining models demonstrated naive performance.

Conclusion

Despite evaluating several state-of-the-art open-source VLMs, none demonstrated strong performance. MedGemma, the smallest model tested, performed best, suggesting domain-specific pretraining may be more important than model size. Merlin produced similar results but accepts only a single CT series, whereas GI bleeding is typically diagnosed using multiphase CT. These findings suggest current VLM architectures and training strategies remain inadequate for radiologic tasks requiring reasoning across multiple slices and imaging phases.

Statement of Impact

Current state-of-the-art foundational vision-language models demonstrate limited performance in detecting gastrointestinal contrast extravasation on CT.

Model	AUC	Optimal Threshold	Sensitivity	Specificity	PPV	F1	Accuracy
MedGemma 1.5 4B	0.592	0.502	0.524	0.632	0.587	0.554	0.578
Llama-3.2-Vision 90B	0.588	0.119	0.678	0.466	0.559	0.613	0.572
InternVL3 78B	0.514	0.148	0.772	0.276	0.516	0.619	0.524
LLaVA-Med v1.5 7B	0.508	0.180	0.868	0.176	0.513	0.645	0.522
Qwen2.5-VL 72B	0.474	0.031	0.070	0.944	0.556	0.124	0.507
Gemma-3 27B	0.467	<0.001	0.892	0.114	0.502	0.642	0.503

Model performance along with metrics

Keywords
Foundation Model; VLM; Artificial intelligence