Evaluation of a Fine-tuned Vision-Language Model for Automatic Generation of Lumbar Spine MRI Reports

Mattia Perrone, John T. Martin
Rush University Medical Center, Chicago, IL,
Email of Presenting Author: mattia_x_perrone@rush.edu

DISCLOSURES: M.P., J.T.M. (N)

INTRODUCTION: Radiologists read clinical musculoskeletal images, identify pathologies and generate corresponding reports at a demanding rate, where both error rate[1] and radiologist burnout[2] increase with the volume of work. Deep learning models with architectures for image processing and generating text have been developed to caption images[3] and more recently to generate radiology reports for clinical imaging tasks[4]. Vision–language models (VLMs) have demonstrated report-generation feasibility in spine imaging, yet prior work relied on small datasets and limited models[4]. For lumbar spine MRI, radiology reports typically consist of a detailed level-by-level analysis of spinal pathology as well as an impression, e.g., a final summary of the patient's overall disease burden. The aim of this study is to implement a VLM built on state-of-the-art architecture that automatically generates structured radiology impressions from lumbar spine MRI. Here, we demonstrate proof-of-concept using paired MRI and radiology reports from our medical center health record.

METHODS: Dataset: Data (IRB approved) were retrieved from the Rush EHR by querying patients who underwent lumbar spine MRI (CPT code: 72148), covering the period from 04/2014 to 03/2024 and lived in Chicago, IL. The search identified 4589 patients and 2100 were included in this initial pilot analysis (age: 56±16 years; 36% male, 64% female). Text preprocessing: Radiology impressions were cleaned by removing all identifiers (e.g. patient information, radiologist name, date of the exam, etc.) and preprocessed via ChatGPT 4o API. ChatGPT prompts were designed to extract, normalize and structure lumbar MRI Impression text, restricting to L4-L5 and L5-S1 and outputting level-specific findings plus a concise clinical summary of spine pathology. Computational workflow (**Fig. 1A**): 2100 T2-weighted sagittal MRI volumes were preprocessed by adjusting spatial resolution to a target of 0.9x0.9x3.5 mm, center padding/cropping to a target shape of 320x320x32 voxels, and voxel intensity normalization. As proof-of-concept, we trained a VLM to read a single mid-sagittal MRI slice and generate a structured results for L4-L5, L5-S1 and an overall impression. We fine-tuned Qwen-2.5 7B[5] using LoRA (r=16, α=32) to optimize attention and MLP weights of the language model component while fixing vision encoder weights. The dataset split was 80/10/10 between training, validation and testing. We compared the fine-tuned model with zero-shot Qwen 2.5 and zero-shot ChatGPT 4 Turbo across three performance metrics for lexical similarity (ROUGE-L), semantic similarity (BERTScore), and a custom F1 score for measuring accuracy of identifying spine pathology (Pathology-F1). Pathology-F1 measures accuracy in predicting spinal pathology presence/absence at each spinal level. Five pathologies were considered in this initial analysis (disc herniation, disc bulging, central canal stenosis, listhesis and fracture) and labels were extracted from the report using pattern-matching methods that allow to automatically detect specific keywords or phrases in text (regex rules). All experiments were performed on NVIDIA A5000 (24GB VRAM).

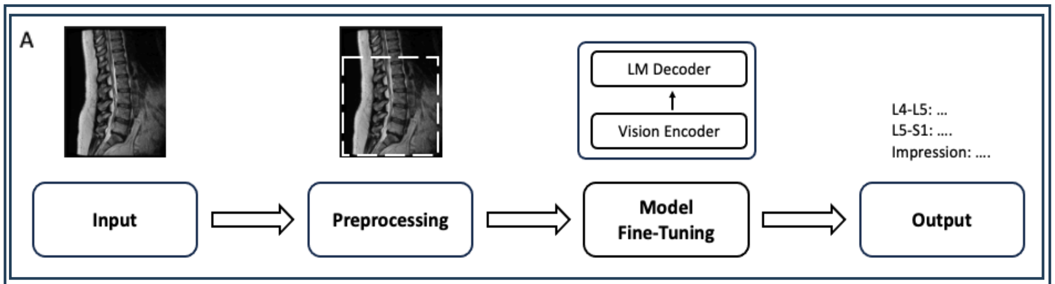
RESULTS: The fine-tuned Qwen model achieved ROUGE-L of 0.437, BERTScore of 0.928. and Pathology-F1 of 0.336 (mean inference time = 7s), outperforming the zero-shot Qwen 2.5 7B and ChatGPT 4o-Turbo baseline (**Fig. 1B**). The Qwen model produced realistic reports with moderate accuracy in identifying spine pathology. In one case (Fig. 1C), fine-tuned Qwen correctly identified listhesis at L4-L5, while missing mild disc degeneration at L5-S1.

DISCUSSION: Towards reducing workloads in musculoskeletal radiology, we generated structured radiology impressions from lumbar spine MRI. We demonstrate preliminary feasibility of this approach, where Qwen 2.5 7B was fine-tuned to read and report on lumbar spine MRI using a single sagittal slice from lumbar spine MRI and an abbreviated radiology report. We found that fine tuning improved performance in comparison to zero-shot Qwen and zero-shot GPT-4 Turbo. Next steps include incorporation of loss functions for pathology-aware training, integration of images from 3D MRI volumes including coronal views, and expansion to the full dataset of 4589 patients.

SIGNIFICANCE: Automated MRI report generation can reduce radiology workloads while maintaining diagnostic accuracy for key spinal pathologies.

REFERENCES: 1) Kasalak et al., 2023, *Eur J Radiol Open*; 2) Fawzy et al., 2023, *Eur J Radiol Open.* 3) Vinyals et al., 2014, *arXiv* ; 4) Yeasin et al., 2024, *Journal of Clinical Medicine,* 5) Yang et al., 2024, *arXiv;*

ACKNOWLEDGEMENTS: NIH R00AR077685



B

Model	ROUGE-L F1 (Lexical)	BERTScore F1 (Semantic)	Pathology F1 (Clinical)
GPT-4 Turbo (zero-shot)	0.272	0.886	0.203
Qwen 2.5 7B (zero-shot)	0.354	0.906	0.200
Qwen 2.5 7B (fine-tuned)	**0.437 (+23 %)**	**0.928 (+2.4 %)**	**0.336 (+68 %)**

C **Example of generated report (Test set)**

LEVEL-SPECIFIC FINDINGS: L4-L5: Grade 1 anterolisthesis with disc space narrowing - L5-S1: None
CLINICAL IMPRESSION: Grade 1 spondylolisthesis of L4 on L5 with associated degenerative changes.

Correspondent ground truth

LEVEL-SPECIFIC FINDINGS: L4-L5: Grade 1 anterolisthesis with moderate disc space narrowing - L5-S1: Mild disc degeneration
CLINICAL IMPRESSION: Grade 1 spondylolisthesis of L4 on L5 with moderate degenerative disc changes. Mild degenerative changes at L5-S1

Figure 1: (**A**) Computational pipeline. (**B**) Comparison of metrics among the three models tested in the study. Performance gains of the fine-tuned model are reported relative to the Qwen 2.5 7B zero-shot baseline. (**C**) Examples of a radiology impressions generated by fine-tuned Qwen 2.5 7B and its correspondent ground truth.