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

SpineGraph: Vertebral-Level-Aware Evaluation for Automated Cervical Spine Report Generation

Sydney A. Levin, Tulane University; Matheus Ferreira; Eduardo Farina, MD; Ronald Peshock, MD; Paulo Kuriki, MD

Introduction/Background

Vision-language report drafters are limited less by text generation than by measurement: without a clinical-correctness metric they cannot be objectively evaluated. Such metrics are predominantly chest-specific. Spine reporting depends on vertebral-level attribution, a documented and frequently undetected error current metrics do not verify. Because the spine is one anatomical continuum, a single level-aware framework can evaluate all of it. We hypothesized granular, level-aware scoring could supply this signal, and introduce SpineGraph, beginning with the cervical spine.

Methods/Intervention

SpineGraph comprises a cervical spine ontology of 53 granular findings under 18 parent categories, separating clinically opposed concepts such as anterolisthesis versus retrolisthesis; a 0.6-billion-parameter model fine-tuned with LoRA on 23,625 reports automatically labeled by GPT-OSS-120B to convert the FINDINGS section into strict-JSON observations with evidence spans, status, and linked locations; and a hierarchical metric over leaf, parent, leaf-location, and parent-location F1, scoring level ranges by vertebral-atom overlap. One senior neuroradiologist reviewed annotations for 100 held-out reports to form the reference. We evaluated the trained model, applied concept and level perturbations, and re-ran the frozen extractor on reports with edited location text. Intervals used 10,000-replicate case bootstrap.

Results/Outcome

The trained model achieved leaf F1 0.919 (95% CI 0.888-0.944), parent F1 0.961, leaf-location F1 0.824 (0.769-0.874), and hierarchical F1 0.906 (0.875-0.933); whole-graph agreement was exact in 55 reports. Same-parent concept substitutions (n=89) kept parent F1 at 1.000 in every case while hierarchical F1 fell 0.154 (0.132-0.180). Level shifts (n=66) reduced hierarchical F1 more when non-overlapping (0.100) than adjacent (0.043). In paired re-inference (70 pairs, no strict-JSON failures), the extractor followed the intended level change in all 63 reports correct at baseline, preserving finding identity. Reviewed weak annotations scored leaf F1 0.985, reflecting label quality, not model performance.

Conclusion

SpineGraph adds the vertebral-level attribution broad-category scoring lacks. Its ontology, extractor, and metric are region-agnostic by construction: the cervical spine shown here is the first stage of a progression through the thoracic and lumbar spine toward a whole-spine standard.

Statement of Impact

By providing a foundational clinical-correctness signal where none existed, SpineGraph enables

benchmarking, quality assurance, and reward-based optimization of generated reports; the blueprint extends to other underserved body regions.

Figure 1. Clinical-correctness metrics for radiology report generation

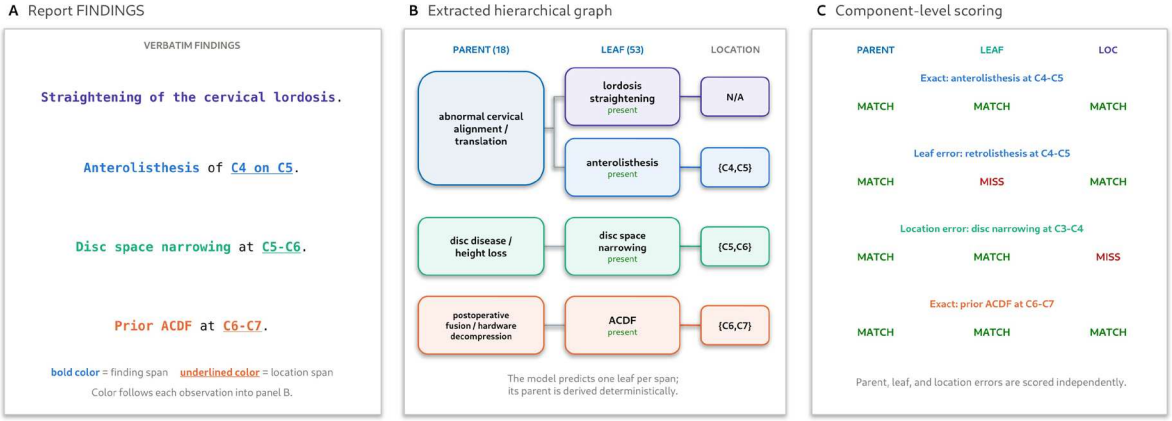
Metric	Target region	Finding granularity	Vertebral level verified	Partial credit	Validation basis
CheXbert F1	Chest	14 thoracic labels	no	none	expert-labelled corpus
RadGraph F1	Chest	entities and relations	no	entity / relation overlap	expert annotation
RadCliQ	Chest	composite of four metrics	no	none	radiologist alignment
GREEN	Chest	LLM error notation	no	error counts	expert error counts
SRR-BERT	Chest	55 labels, hierarchical	no	taxonomy-based	5-radiologist reader study
SpineGraph (this work)	Cervical spine	53 leaves / 18 parents	yes	hierarchical and level-overlap	clinician-corrected reference, perturbation, paired re-inference

Partial credit = whether a clinically related but non-identical prediction receives graded rather than all-or-nothing credit.

Established clinical-correctness metrics for report generation target the chest. **None of the metrics listed verifies the vertebral level to which a finding is attributed**; RadGraph encodes anatomical entities but is chest-oriented and does not check vertebral-level attribution. Vertebral-level discrepancies are a documented and frequently uncorrected error class in spine reporting, so this axis is unmeasured by current report-generation metrics. SpineGraph adds level-indexed scoring for cervical spine radiograph reports.

Fig 1. Clinical-correctness metrics for radiology report generation

Figure 2. Representative extraction and hierarchical scoring



Hierarchical F1 = 0.50 · leaf F1 + 0.20 · parent F1 + 0.20 · leaf-location F1 + 0.10 · parent-location F1

Location ranges are expanded into canonical vertebral atoms (C5-C7 → {C5, C6, C7}) and scored by set overlap, so an overlapping level retains partial credit and a non-overlapping level does not.

Broad category scoring can give full credit to same-parent substitutions and cannot assess shifted vertebral levels; the hierarchical metric penalizes leaf identity and level attribution separately.

Fig 2. Representative extraction and hierarchical graph scoring

Figure 3. Validation results

A Extraction performance against the clinician-corrected reference (N = 100 reports)

Metric	Trained SpineGraph	95% CI	GPT-OSS 120B weak annotations (label quality, not model performance)	95% CI
Leaf F1	**0.919**	0.888–0.944	0.985	0.975–0.994
Parent F1	0.961	0.935–0.979	0.995	0.988–1.000
Leaf-location F1	0.824	0.769–0.874	0.963	0.940–0.983
Parent-location F1	0.893	0.850–0.931	0.982	0.964–0.996
Hierarchical F1	**0.906**	0.875–0.933	0.982	0.972–0.991

Exact agreement (trained model): active leaf set 0.59, parent set 0.77, whole hierarchical graph 0.55. Location metrics computed on the 96 reports containing at least one location. Intervals are 10,000-replicate case bootstrap.

B Metric response to controlled perturbation of the reference graphs

Perturbation	n	Parent F1	Reduction in hierarchical F1	95% CI
Same-parent concept substitution	89	unchanged (1.000)	**0.154**	0.132–0.180
Vertebral level shift, overlapping range	66	unchanged (1.000)	**0.043**	0.036–0.052
Vertebral level shift, non-overlapping range	66	unchanged (1.000)	**0.100**	0.082–0.120

C Paired end-to-end re-inference

Outcome	Value
Report pairs evaluated	70
Strict-JSON extraction failures	0
Target observation correct at baseline	63 / 70
Intended level followed (baseline-correct)	**63 / 63**
Finding identity preserved	63 / 63
Non-target leaf labels/statuses unchanged	61 / 63

(A) Trained-model predictions and the underlying GPT-OSS 120B weak annotations, each scored against the same clinician-corrected reference; the weak-annotation column measures label quality and is not a model result.

(B) Reference graphs were deterministically perturbed and rescored without model inference. Same-parent substitutions preserve parent F1 while changing leaf identity; level shifts preserve both concept components.

(C) One location span was edited in the report and the frozen extractor re-run, conditioned on baseline-correct targets. Non-target leaf labels/statuses were unchanged in 61/63. Strict graph agreement was 51/63 because 10 pairs had expected co-referenced location changes; 2/63 showed other leaf/count changes. Inference used vLLM with temperature 0 and JSON-constrained output.

Fig 3. Validation results

Keywords

cervical spine**;** radiology report; vision-language model; graph evaluation; ontology; information extraction