



From Narrative Radiology Reports to Vision-Language Model Training Labels: Transparent Physician Validation Using Evidence-Grounded Visualization

Mohammed Ayman Habib, PhD Student, University of Utah; Isaac Anthony; Osei Brempong, PhD Student; Derrick Wong; Morteza Fayazi, PhD; Seyyedkazem Hashemizadehkolowri, PhD; John Roberts, PhD; Tyler Richards, MD; Maryam Soltanolkotabi, MD

Introduction/Background

We developed an evidence-grounded large language model (LLM) pipeline to convert more than 33,000 scoliosis radiology reports into structured VLM training labels. Because physician validation is the principal scalability bottleneck, each extracted variable is displayed alongside color-coded highlights of the supporting report text. We evaluated the usability of this interface.

Methods/Intervention

An evidence-grounded LLM extracted structured scoliosis variables including Cobb angles, curve characteristics, spinal hardware, coronal and sagittal imbalance, vertebral numbering, compression fractures, and incidental findings. The interface simultaneously displayed the structured extraction and color-coded/highlighted the evidence within the radiology report text that was used to derive each extracted variable. Two physician reviewers independently validated the first 100 reports using a randomized crossover design in which each report was reviewed once using the evidence-grounded interface and once using the original report interface. Review time, reviewer perception, inter-reviewer agreement, and field-level correction patterns were assessed.

Results/Outcome

The first 100 reports contained a mean of 6.3 structured variables per report with balanced complexity between arms. Mean review time was comparable for evidence-highlighted and original reports (Reviewer 1: 24.3 vs. 24.3 min per 10-report block; Reviewer 2: 21.0 vs. 22.8 min) as was validation performance across extraction fields. Both reviewers were faster over the final 20 reports (19% and 46%) suggesting experience contributed more to review time than the interface. Inter-reviewer agreement was 92.2% across 1,052 field-level comparisons. The interface was rated helpful in 78% of evaluations, neutral in 12%, and slightly confusing in 10%.

Conclusion

In this pilot study, evidence-grounded visualization maintained comparable review times, high inter-reviewer agreement, and similar validation performance. Although it did not improve validation speed or accuracy, reviewers consistently reported perceived improved clarity and easier navigation between structured outputs and their supporting text. The primary benefit therefore appears to be improved usability and sustainability of physician validation workflows rather than faster review. Greater efficiency gains may emerge for longer

reports in which locating sparse supporting evidence is more time-consuming.

Statement of Impact

Color-coded evidence grounding of LLM-extracted variables showed potential to reduce the cognitive burden of physician validation, a principal bottleneck in scaling structured data extraction from radiology reports.

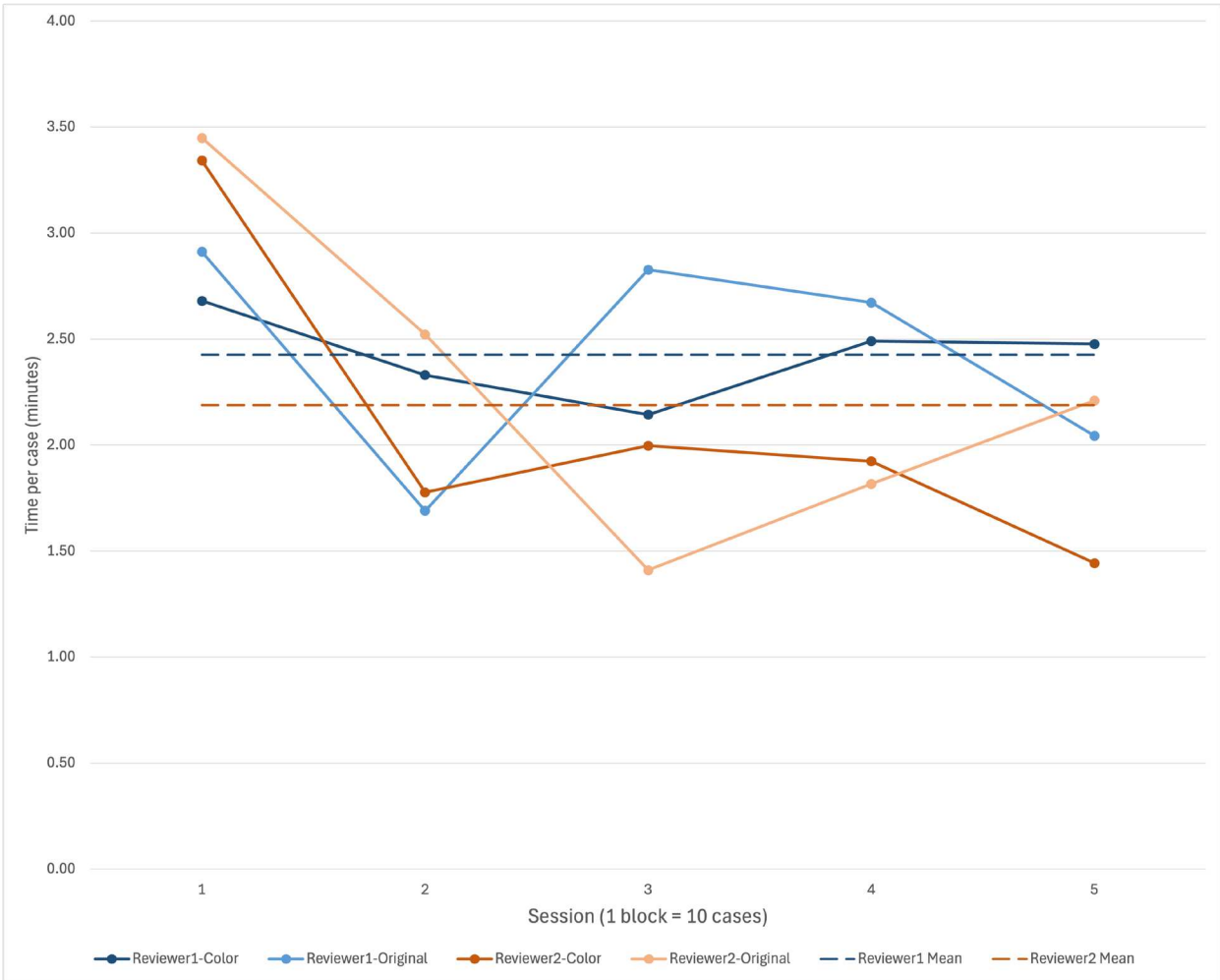


Figure 1. Impact of color-coded reports compared to original reports on time to complete the review of 10 cases across two reviewers over time.

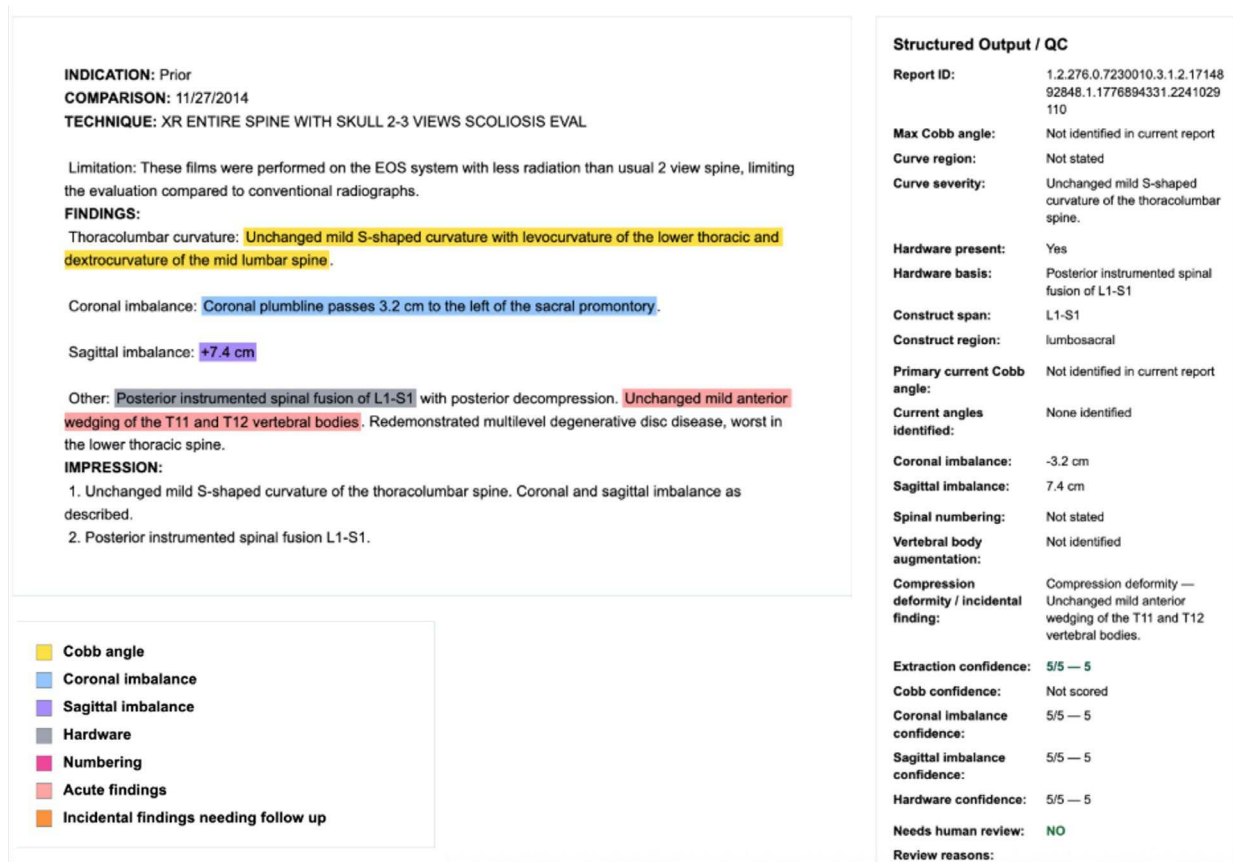


Figure 2. Example of a color-coded report to demonstrate the evidence used by the LLM for data field extraction. The key for the color-coding is located in the bottom left corner and the normalized extracted variables are listed on the right.

Rating	Reviewer 1, n (%)	Reviewer 2, n (%)	Combined, n (%)
Very helpful	5 (10%)	30 (60%)	35 (35%)
Somewhat helpful	30 (60%)	13 (26%)	43 (43%)
Neutral	11 (22%)	1 (2%)	12 (12%)
Slightly confusing	4 (8%)	6 (12%)	10 (10%)
Helpful (Very + Somewhat)	35/50 (70%)	43/50 (86%)	78/100 (78%)

Table 1. Summary of reviewer ratings of the usefulness of the color-coding of reports to verify the data extracted by the large language model. Of note, none of the color-coded reports were rated as "not useful".

Keywords

Large Language Model; Data extraction; Validation