



Personalizing Radiology Impression Generation to Radiologist Style and Subspecialty with Auditable, Auto-Generated Playbooks

Ushanandini Raghavan, PhD, DeepHealth; Madhu Jahagirdar

Introduction/Background

Generating impressions using Language Models (Large and Small) is established practice. Many studies focus on model generalizability, but little attention has been paid to personalizing AI-generated impressions to a specific radiologist's style and subspecialty. Our contribution enhances an existing impression-generating LLM's system prompt with a scope-specific Playbook: a modular, human-readable set of rules, surfaced to the radiologist for review, guiding how findings are ordered and presented in the impression.

Methods/Intervention

We used the recently proposed Agentic Context Engineering (ACE) method and adapted it to generate Playbooks. Unlike prompt optimization or fine-tuning, where enhancements stay hidden in prompts or weights, ACE produces an explicit, reviewable Playbook. ACE works offline, reflecting on differences between AI-generated and ground-truth impressions along two axes: ordering (which finding types were prioritized, deprioritized, or omitted) and presentation (whether findings were consolidated or enumerated, and what was emphasized versus deferred). Reflections are curated into the Playbook. We measured lexical similarity (bigram Jaccard, rouge-l-f1), semantic similarity (BERTScore), and structural similarity (word count) against a no-Playbook baseline. An LLM judge assessed whether characterizations in the generated impression conformed to the input findings.

Results/Outcome

We applied ACE across four modalities (XR, CT, MRI, US) and six scan types (chest, neck, abdomen, lumbar spine, knee, shoulder) from 4 radiologists, with 300 training and 100 eval reports per scope. Lexical, semantic, and structural metrics improved with the Playbook over baseline in every dataset. The LLM judge showed a slight uptick in non-conforming characterizations under the Playbook, ranging 0–5% across datasets. The generated Playbook was a consistent set of rules, for example, the XR lumbar spine Playbook had 32 rules total, with callouts such as "Ordering: on flexion-extension studies, lead with/close on the dynamic instability assessment" and "Presentation: consolidate degenerative findings into one statement, omit level-by-level grading."

Conclusion

We auto generated modular Playbooks that can be reviewed by a radiologist, reducing friction in personalizing radiology impressions for practice.

Statement of Impact

Surfacing personalization as a reviewable Playbook, rather than a hidden prompt or fine-tuned weights, gives radiologists direct visibility and control over how AI adapts to their style, lowering the barrier to individualized adoption of impression-generation tools in clinical practice.

ACE generated Playbook for XR lumbar spine of radiologist 55

All 32 rules below were auto generated by ACE's Reflector/Curator from ground-truth reports (backfilled data) — not manually authored. Organized into two axes: Ordering (which finding types are prioritized, deprioritized, or omitted) and Presentation (how findings are consolidated, phrased, and emphasized). Surfaced to the radiologist for review, edit, and acceptance before use.

Ordering (18 rules)

1. Lead with the primary positive finding most directly relevant to the clinical indication; deprioritize or omit incidental, secondary, and normal findings entirely, deferring their detail to the report body.

2. Treat the absence of acute osseous findings as a secondary qualifier appended to the primary positive finding, not as a lead or standalone item; omit entirely when the impression is dominated by chronic degenerative findings that implicitly exclude acute pathology.
3. When the exam is globally normal, collapse the entire impression into a single overarching normal statement rather than enumerating individual structures or negative findings.
4. When a postoperative or surgical history finding is present and clinically relevant, surface it as a distinct high-priority finding, equal to or above associated degenerative findings.
5. When the exam is globally normal, omit explicit negative statements for specific categories (acute osseous abnormality, fracture, dynamic instability); cover the study with one summary normal statement.
6. When degenerative findings are mild or incidental relative to the clinical question, suppress them entirely (facet arthrosis, scoliosis, transitional vertebrae, vacuum disc phenomenon) rather than listing as secondary items.
7. On flexion-extension or bending studies, lead with or close on the functional assessment of dynamic instability as the primary exam-specific answer, regardless of degenerative findings present.
8. When the exam is globally normal, prefer a single summary of absence of acute abnormality ('no acute abnormality seen', 'unremarkable radiographs') over an affirmative declaration of complete normalcy.
9. When mild chronic findings accompany an otherwise normal exam, surface the chronic finding briefly before a trailing negative qualifier; suppress entirely only when incidental relative to the clinical question.
10. Omit incidental extraskeletal or non-spinal findings (abdominal calcifications, unrelated hardware) from the impression entirely.
11. Apply a binary threshold for incidental chronic findings alongside a normal exam: either surface as sole item with trailing qualifier, or suppress entirely — never enumerate both as co-equal items.
12. When multiple degenerative finding types (disc space narrowing, spondylolisthesis, facet arthrosis) are all present, treat as a unified complex; lead with the most salient and fold others into a subordinate clause.
13. When the impression is a single global normal conclusion, prefer a bare unnumbered statement over an explicit negative for a specific finding category.
14. When a clinical recommendation (e.g., flexion-extension views) is warranted by a positive finding, append it as a trailing clause to that finding rather than a separate item.
15. When reporting a postoperative/arthrodesis status exam, anchor to overall surgical stability and interval change rather than enumerating hardware components.
16. When globally normal, prefer 'unremarkable radiographs of the lumbar spine' over 'no acute abnormality seen' as default; reserve acute-anchored phrasing for trauma or rule-out contexts.
17. When a compression fracture is present alongside incidental degenerative findings, always lead with the fracture regardless of acuity; relegate degenerative findings to a subordinate statement.
18. When globally normal on bending/flexion-extension views, omit an explicit dynamic-instability negative; cover it implicitly within the single global normal summary.

Presentation (14 rules)

19. Consolidate related degenerative findings (disc space narrowing, facet arthrosis, spondylolisthesis) into a single unified statement; omit level-by-level severity grading in favor of a broad categorical descriptor.
20. Omit explicit statements confirming normal structures (maintained vertebral heights, preserved disc spaces, absence of soft tissue abnormality); cover normality implicitly rather than listing each item.
21. Use a prose or minimally structured format with few or no numbered items; separate multiple findings by paragraph or semicolon rather than discrete numbered lines.
22. When acuity/chronicity is uncertain, integrate an explicit qualifier (e.g., 'favor chronic but technically indeterminate') directly into the finding statement.
23. Omit severity grading qualifiers (mild, moderate, severe) from impression-level statements when summarizing degenerative findings; reserve severity detail for the report body.
24. When a quantitative measurement (e.g., mm of listhesis) is available, substitute it directly for a qualitative grade rather than pairing both.
25. When summarizing a globally normal study, anchor the statement to the radiographic study itself ('unremarkable radiographs of the lumbar spine') rather than the anatomical structure alone, as a single unformatted prose statement.
26. When reporting scoliosis, omit quantitative degree measurements; characterize by type and location, integrated with associated degenerative findings into one consolidated statement.
27. When referencing absence of dynamic instability, integrate it as a closing clause rather than a discrete numbered line.
28. When consolidating multiple degenerative types, prefer a broad categorical label as anchor, then append specific subtypes as subordinate descriptors within the same clause.
29. When consolidating degenerative scoliosis with associated disc/facet findings, treat scoliosis as causally integrated rather than a separately enumerated finding.

- 30. When the impression is globally normal, present as a single bare unnumbered prose statement without numbered structure or bold headers.
- 31. When reporting a compression fracture with established/favored chronicity, embed the chronicity characterization directly within the fracture description rather than as a standalone qualifier.
- 32. When the impression is a single chronic finding with no acute pathology, let it stand alone without an added negative acute clause.

modality	scan_type	radiologist	Words	Words_Base	Words_J	BERT_F1 ^Δ	BERT_F1 ^Δ	BERT_F1 ^Δ	Bigram_Jaccard ^Δ	Bigram_Jaccard ^Δ	Bigram_Jaccard ^Δ	ROUGE_L ^Δ	ROUGE_L ^Δ	ROUGE_L ^Δ	Clinical_Support ^Δ	Clinical_Contradict ^Δ	Clinical_Support ^Δ	Clinical_Contradict ^Δ	Clinical_Contradict ^Δ
						eline	CE	delta	baseline	CE	delta	Baseline	ACE	delta	Pct_Baseline	Pct_Baseline	Pct_ACE	Pct_ACE	delta
CT	abdomen		36	41.7	71.8	53.4	0.903	0.917	0.014	0.168	0.182	0.014	0.397	0.453	0.056	99	1	97	3
CT	abdomen		55	16.5	56	30.8	0.858	0.889	0.031	0.036	0.11	0.074	0.158	0.305	0.147	100	0	97	3
CT	chest		55	19.9	74.1	46.9	0.837	0.861	0.024	0.027	0.04	0.013	0.129	0.177	0.048	100	0	100	0
MRI	knee		86	26.7	89.9	42.6	0.879	0.917	0.038	0.103	0.183	0.08	0.277	0.465	0.188	100	0	100	0
MRI	lumbar_spine		36	40.6	159	89.7	0.871	0.888	0.017	0.094	0.098	0.004	0.246	0.316	0.07	100	0	98	2
MRI	shoulder		86	35.9	50.8	50.4	0.918	0.933	0.015	0.196	0.243	0.047	0.39	0.477	0.087	100	0	100	0
US	abdomen		37	22.6	66.5	36.3	0.852	0.879	0.027	0.031	0.096	0.065	0.136	0.261	0.125	100	0	99	1
US	neck		55	158.2	60.4	114.3	0.805	0.842	0.037	0.008	0.034	0.025	0.134	0.326	0.192	96	4	95	5
XR	chest		36	8.6	31.8	10.1	0.857	0.949	0.092	0.016	0.435	0.419	0.122	0.623	0.501	99	1	99	1
XR	chest		55	5.7	22	8.7	0.847	0.959	0.112	0.037	0.588	0.551	0.085	0.744	0.659	100	0	99	1
XR	lumbar_spine		55	8.3	20.5	17.5	0.873	0.908	0.035	0.053	0.287	0.234	0.239	0.428	0.189	100	0	97	3
XR	shoulder		55	7	23.1	6.2	0.869	0.905	0.036	0.028	0.086	0.058	0.184	0.303	0.119	100	0	100	0

Fig. 1 Results for data sets comparing playbook vs baseline on structural, lexical, semantic and LLM judge metrics

Keywords

agentic context engineering; radiology impressions; personalization; playbook