Automating AO/OTA Coding from Radiology Reports and Resident Descriptions Using Retrieval-Augmented Large Language Models

Daniel Theng[1], Timothy Keeley[1], Timothy Wu[1], Meir Marmor[1]
[1]University of California, San Francisco,
Daniel.Theng@ucsf.edu

Disclosures: None

INTRODUCTION: Large Language Models (LLM) have potential in enhancing medical documentation. Due to their broad and extensive training data, LLMs have demonstrated strengths in general knowledge, but often struggle to adapt to more specific use cases. Retrieval augmented generation (RAG) addresses these shortcomings by providing LLMs with external data sources that the LLM is required to reference when constructing responses. Therefore, RAG can increase accuracy, transparency, and reliability while reducing hallucinations for problems that require more niche knowledge. One such use case is within fracture classification, which is critical for research, consistent communication between providers, and clinical quality assessment. Although the AO/OTA fracture classification system underpins trauma decision-making, research, and communication, it is rarely included in documentation. Its hierarchical complexity makes coding labor-intensive and variable, with only moderate agreement at finer tiers. This study aims to evaluate whether and academic health system HIPAA-compliant GPT-o3-mini, could effectively classify fractures described in radiology reports and resident descriptions using a RAG approach leveraging the AO/OTA Fracture and Dislocation Compendium as a reference document.

METHODS: 30 radiology reports of radiographs from a level 1 trauma center presented at a fracture conference between the dates of 09/10/2024 to 10/10/2024 were collected. Two independent reviewers extracted radiology report text pertaining to the primary fracture. Members of an OTA Classification Committee classified these according to the AO/OTA classification fractures separately, first based on the extracted radiology reports alone and then using the original x-ray images. The AO/OTA fracture classification compendium was converted into a simplified text document by removing all tables and images, in order to create a reference document for the RAG approach. These extracted reports were used as inputs, and the LLM was queried to retrieve the AO/OTA classification that best describes the extracted text. The prompt was designed following heuristic prompt engineering strategies such as strict output structure, define model role, and refraining from outputting chain-of-thought. LLM's classifications were compared against these labels, and accuracies (categorical and cumulative) were calculated. A similar protocol was followed, but with resident descriptions. During fracture conference, orthopedic surgery residents describe fractures presented. These descriptions were also collected, and members of the OTA Classification Committee classified these descriptions, and the accuracies (categorical and cumulative) were calculated. Lastly, to establish a "gold standard", the OTA classification Committee was provided the radiographs of the fractures and asked to classify the images.

RESULTS SECTION: All analyses used n = 30 fracture cases.
LLM (zero-shot) on radiology reports vs Committee on reports.
Accuracy by category (correct/total, %): Bone 26/30 (86.7%), Location 25/30 (83.3%), Type 19/30 (63.3%; 95% CI 45.5–78.1), Group 13/30 (43.3%; 95% CI 27.4–60.8), Subgroup 25/30 (83.3%), Qualifications 23/30 (76.7%), Universal modifiers 12/30 (40.0%). Cumulative "all elements" 4/30 (13.3%; 95% CI 5.3–29.7).

LLM (zero-shot) on resident descriptions vs Committee on resident descriptions.
Bone 24/30 (80.0%), Location 22/30 (73.3%), Type 12/30 (40.0%; 95% CI 24.6–57.7), Group 12/30 (40.0%; 95% CI 24.6–57.7), Subgroup 26/30 (86.7%), Qualifications 23/30 (76.7%), Universal modifiers 21/30 (70.0%). Cumulative "all elements" 3/30 (10.0%; 95% CI 3.5–25.6).

LLM on radiology reports vs Committee on radiograph images.
Bone 24/30 (80.0%), Location 23/30 (76.7%), Type 19/30 (63.3%; 95% CI 45.5–78.1), Group 14/30 (46.7%; 95% CI 30.2–63.9), Subgroup 17/30 (56.7%), Qualifications 19/30 (63.3%), Universal modifiers 11/30 (36.7%). Cumulative "all elements" 3/30 (10.0%; 95% CI 3.5–25.6).

Committee on radiology reports vs Committee on radiograph images.
Bone 27/30 (90.0%), Location 24/30 (80.0%), Type 17/30 (56.7%; 95% CI 39.2–72.6), Group 15/30 (50.0%; 95% CI 33.2–66.8), Subgroup 20/30 (66.7%), Qualifications 23/30 (76.7%), Universal modifiers 9/30 (30.0%). Cumulative "all elements" 3/30 (10.0%; 95% CI 3.5–25.6).

DISCUSSION: We evaluated whether a zero-shot, retrieval-augmented LLM can assign AO/OTA fracture classifications from narrative text (radiology reports; resident descriptions) compared with expert committee labels from text and from radiographs. The LLM recovered upper-level AO/OTA elements well (Bone/Location ≥73–87%) and Type ~60–63%; performance dropped at Group and for Universal modifiers, paralleling the pattern seen when humans classify from text alone. Against radiograph-based committee labels, LLM (text) remained comparable at Type and within range for Group/Subgroup, while trailing for Bone, indicating that most, but not all, discriminative signal is present in radiology reports/resident description.

Given the high accuracies for Bone/Location, a workflow tool could utilize an LLM to pre-populate less granular fields of the classification from radiology reports or resident descriptions, allowing for further effort towards aspects of the classification that are more contentious. A practical workflow is to auto-populate Bone/Location/Type from reports or resident descriptions and route Group/Subgroup/modifiers for expert verification. This preserves speed gains while concentrating expert effort where agreement is hardest. Moreover, inclusion of resident description data provides utility in identifying trainee gaps that may require further attention. This way it is possible to use an LLM as an educational tool to accelerate trainee learning.

The lower tiers lag in performance could be attributed to decreased rates of documentation. Classification of subclasses and modifiers depend on granular language and context that are inconsistently documented in text, leading to steep cumulative-accuracy decay once Group and modifiers are required.

Study is limited by a single center, n=30, one-month sampling; zero-shot single-model configuration; no inter-rater reliability estimates reported for committee labels in this dataset; radiographs treated as the reference without adjudication details. As with all clinical NLP tools, they are subject to hallucinations and thus require expert verification prior to clinical use. These factors likely cap ceiling performance at granular tiers. Future work should emphasize increasing sample sizes and sites, and trial multi-shot prompting strategies targeted to Group/Subgroup decision points. Lastly, RAG architecture to utilize richer/well-structured sources, such as decision trees, may improve precision and reduce hallucinations.

SIGNIFICANCE/CLINICAL RELEVANCE: AO/OTA fracture coding is vital for communication between providers but remains labor-intensive and variable among providers. This study aims to explore the feasibility of a zero-shot retrieval-augmented LLM that reliably labels AO/OTA codes from routine radiology reports or resident descriptions, advancing technical capabilities and ultimately improving provider workflow.