

Natural Language Processing and Large Language Model-Based Automated AO/OTA Classification of Proximal Femoral Fractures: A MIMIC-IV Retrospective Study

Yuhang Jin

Shandong University of Traditional Chinese Medicine

Zhipeng Fan

The 960th Hospital of the PLA Joint Logistics Support Force

Xuesong Yan

The 960th Hospital of the PLA Joint Logistics Support Force

Jiawei Li

The 960th Hospital of the PLA Joint Logistics Support Force

Zhuohang Wu

The 960th Hospital of the PLA Joint Logistics Support Force

Yongxian Zhang

Zhangyongxian513@126.com

The 960th Hospital of the PLA Joint Logistics Support Force

Research Article

Keywords: Retrieval-Augmented Generation, Large Language Models, AO/OTA Classification, Proximal Femoral Fractures, MIMIC-IV, Natural Language Processing, Clinical Decision Support

Posted Date: September 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10902997/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Accurate AO/OTA classification of proximal femoral fractures is a critical step in surgical decision-making and prognostic evaluation; however, the complexity of the classification system and the heterogeneity of terminology in radiology reports pose significant challenges for large-scale, standardized application of the classification. In recent years, with the advancement of AI technologies, large language models (LLMs) equipped with retrieval-augmented generation (RAG) capabilities have offered new possibilities for automatically extracting fracture classification information from unstructured radiology reports.

Methods

This study conducted a retrospective analysis of 23,179 radiology reports from the MIMIC-IV database. Through a multi-stage annotation process involving independent labeling by two orthopedic surgeons and adjudication by a senior expert, a total of 157 acute cases of AO/OTA type 31A (intertrochanteric fractures) and type 31B (femoral neck fractures) were ultimately included. We evaluated three experimental models: zero-shot prompting, few-shot learning (5 examples per class), and RAG-augmented prompting (incorporating a structured AO classification knowledge base). Classification performance was assessed using accuracy, F1-score, and confusion matrices. All type 31C (femoral head fractures) cases underwent expert re-adjudication to evaluate annotation quality.

Results

The RAG-LLM framework achieved an accuracy of 94.29% (F1 = 94.29%) in distinguishing 31A from 31B fractures on a balanced test set of 70 cases (35 each of 31A and 31B). Zero-shot prompting attained the same accuracy (94.29%, F1 = 94.28%), while few-shot learning unexpectedly dropped to 88.57% (F1 = 88.53%). Confusion matrices showed minimal misclassifications: 2/35 cases of 31A were misclassified as 31B, and 2/35 cases of 31B were misclassified as 31A. A key finding was that after expert re-adjudication of 47 cases initially labeled as 31C, only 6 (12.8%) met the true AO/OTA 31C criteria; the remaining 41 (87.2%) were actually acetabular fractures (AO/OTA type 61) or misclassified 31A/31B cases. The subtype classification for 31B fractures (31B1 vs. 31B2) achieved an accuracy of 70.59%, reflecting the inherent difficulty of text-based subtype differentiation.

Conclusion

The RAG-LLM demonstrated excellent performance in binary classification of AO/OTA 31A versus 31B proximal femoral fractures using real-world radiology reports. The systematic bias found in 31C labeling

within MIMIC-IV highlights the critical importance of expert data quality assurance in real-world evidence studies. The comparable performance between zero-shot and RAG models suggests that modern LLMs already possess inherent knowledge of AO/OTA classification, though RAG may offer advantages in more nuanced tasks. These findings support the potential of LLM-assisted workflows to streamline fracture classification in clinical and research settings.

Introduction

Proximal femoral fractures are among the most common and clinically significant injuries in orthopaedic trauma. With increasing disability, mortality, and healthcare resource consumption associated with this fracture type(1, 2), the AO Foundation and the Orthopaedic Trauma Association (OTA) established a comprehensive fracture classification system that divides proximal femoral fractures into three main types: 31A (intertrochanteric, extra-articular), 31B (femoral neck, intra-articular), and 31C (femoral head) (3)(**Fig. 1**). Accurate and systematic classification of fractures is essential for selecting appropriate surgical interventions, predicting healing trajectories, and prognostic analysis(4, 5).

Figure 1. AO/OTA type 31 classification of proximal femoral fractures.

[Insert Figure 1 here]

Although the AO/OTA classification of proximal femoral fractures holds significant clinical value, consistent application of the AO/OTA system remains challenging. Even for relatively simple fractures, inter-observer agreement is often suboptimal, with reported Cohen's kappa values ranging from 0.53 to 0.79 (5, 6). For more complex fracture patterns, agreement may drop even further(5). In 2018, the revised AO/OTA classification aimed to simplify the categorisation of intertrochanteric fractures; however, inter-observer reliability at the subgroup level remained "moderate" to "fair" (6). These reliability and consistency issues underscore the need for standardised, automated fracture classification methods to complement clinical judgment and reduce variability.

The widespread adoption of electronic health records and the emergence of large-scale clinical databases such as MIMIC-IV have provided unprecedented opportunities for applying artificial intelligence to clinical problems (7, 8). MIMIC-IV contains detailed radiology reports documenting fracture findings in free-text format, offering a rich yet underutilised resource for orthopaedic research(9).

In recent years, the rapid development of large language models has demonstrated remarkable capabilities in understanding and generating clinical text(10, 11). LLMs can extract structured information from unstructured radiology reports with accuracy approaching that of human experts (12, 13). In the field of orthopaedic trauma, preliminary studies have explored the potential of LLMs for fracture classification. ChatGPT has shown the ability to classify intertrochanteric fractures according to AO/OTA criteria(14), and one study evaluated the performance of ChatGPT-4o and other LLMs in AO fracture classification based on CT text reports, finding that ChatGPT-4o achieved the highest overall

accuracy of 74.6% (15). However, LLMs are constrained by their training data and may produce hallucinations or inconsistent outputs when faced with specialised domain tasks (16).

Retrieval-augmented generation (RAG) has emerged as a powerful technique to address these limitations by dynamically retrieving relevant knowledge from external sources during inference(17). RAG has shown promise in various clinical applications, including medical coding, diagnostic classification, and clinical decision support (18, 19). Recent studies indicate that RAG can significantly improve LLM reliability by anchoring responses to trustworthy external sources, and multi-step retrieval frameworks have been proven effective in improving radiology question-answering performance(20–22).

Notably, most existing studies rely on fabricated or single-centre reports, limiting the generalisability of the models (15, 23). To our knowledge, no study has systematically evaluated the performance of RAG-augmented LLMs for AO/OTA proximal femoral fracture classification using real-world data from a large public database, combined with rigorous expert re-adjudication of label quality.

Therefore, this study aims to: 1) develop and validate a RAG-LLM framework for automatic classification of AO/OTA type 31A versus 31B proximal femoral fractures using MIMIC-IV radiology reports; 2) compare the performance of zero-shot, few-shot, and RAG-augmented prompting strategies; 3) evaluate the quality of AO/OTA annotations in the MIMIC-IV dataset through expert re-adjudication; and 4) explore the feasibility of subtype classification within the 31B category.

Methods

1. Study Design and Data Source

This retrospective cohort study utilised the MIMIC-IV database (version 3.1), a publicly available, de-identified electronic health record dataset containing information on patients admitted to Beth Israel Deaconess Medical Center between 2008 and 2019 (24). The MIMIC-IV database offers a large sample size with detailed radiology reports and has been previously validated in orthopaedic research (25). The study was conducted in accordance with the Declaration of Helsinki and the MIMIC-IV data use agreement. **Figure 2** shows the complete workflow of this experiment.

Figure 2. Flowchart of case identification, annotation, and dataset preparation.

[Insert Fig. 2 here]

2. Data Extraction and Initial Screening

We extracted all radiology reports (n = 23,179) from the radiology table of MIMIC-IV and applied structured queries to filter reports containing the keyword "fracture" in conjunction with proximal femur-specific anatomical terms ("femoral neck," "subcapital," "intertrochanteric," "hip fracture"). Reports

explicitly documenting "no fracture" or "without fracture," as well as those without any relevant findings, were excluded. This initial screening yielded 4,034 reports of acute proximal femoral fractures.

3. Annotation and Classification

A two-stage annotation process was employed to establish the gold-standard classification:

Stage 1: Initial labelling. Two independent orthopaedic surgeons (with 5 and 8 years of clinical experience, respectively) reviewed the impression and findings sections of each report and assigned an AO/OTA classification (31A, 31B, or 31C) according to the AO/OTA Fracture and Dislocation Classification Compendium-2018(3). Discrepancies were resolved through discussion, with adjudication by a senior orthopaedic trauma surgeon with 15 years of experience when necessary.

Stage 2: 31C re-adjudication. Given the clinical and anatomical distinctiveness of 31C (femoral head) fractures, all cases initially labelled as 31C underwent a second independent review by the senior surgeon to verify classification accuracy. This process aimed to assess labelling quality and identify potential systematic biases in the initial annotations.

4. Dataset Preparation

From the annotated dataset, we excluded 31C cases as well as those with indeterminate or uncertain classifications. The final analytical dataset comprised 157 confirmed cases of either 31A (n = 78) or 31B (n = 79). Cases were randomly split at the patient level into a training set (n = 87), a validation set (n = 22), and a test set (n = 48), using stratified sampling to maintain class balance across splits. The validation and test sets were subsequently merged into a final evaluation set of 70 cases (31A: 35, 31B: 35) to enhance statistical power for performance assessment.

5. LLM Configuration

We employed DeepSeek-Chat (deepseek-chat) as the base LLM for all experiments (26). DeepSeek-Chat was selected for its demonstrated performance in clinical text classification tasks, cost-effectiveness, and accessibility to researchers without requiring substantial computational resources. The model was accessed via the DeepSeek API, with the temperature parameter set to 0.1 to minimise output variability and maximise reproducibility.

6. Prompt Engineering and Experimental Conditions

Three experimental conditions were evaluated:

Zero-shot prompting. The model was provided with the impression and findings sections of each radiology report and instructed to output a classification result (31A or 31B) in JSON format, along with its reasoning and confidence assessment. No examples or external knowledge were supplied.

Few-shot prompting. In addition to the report information, the model received 5 representative examples per class (10 examples in total) selected from the training set. Examples were chosen based on impression length and clarity to maximise instructional value.

RAG-augmented prompting. The model received the report information, few-shot examples, and a structured AO/OTA knowledge base summarising the key classification criteria for 31A and 31B fractures. The knowledge base was manually curated based on the AO/OTA Compendium-2018 and formatted as plain text for direct integration into the prompt.

7. Outcome Measures

The primary outcome measure was classification accuracy in distinguishing 31A from 31B fractures. Secondary outcome measures included F1-score, precision, recall, and confusion matrix analysis. In the subtype analysis, we evaluated the accuracy of distinguishing 31B1 (subcapital/non-displaced) from 31B2 (transcervical) fractures.

8. Statistical Analysis

All analyses were performed using Python (version 3.13) and the scikit-learn library. Accuracy was calculated as the proportion of correctly classified cases. The F1-score was computed as the harmonic mean of precision and recall, using weighted averaging. Confusion matrices were generated to visualise misclassification patterns. Given the exploratory nature of this study and the moderate sample size, no formal hypothesis testing was conducted; instead, we report descriptive performance metrics with 95% confidence intervals where applicable.

Results

From the initial 23,179 radiology reports, we identified 4,034 cases of acute proximal femoral fractures. After applying exclusion criteria and confirming classifications, the final analytical dataset comprised 157 cases: 78 (49.7%) classified as 31A (intertrochanteric fractures) and 79 (50.2%) as 31B (femoral neck fractures). The splits into training ($n = 87$), validation ($n = 22$), and merged test ($n = 70$) sets maintained class balance, with 35 cases of each type in the final evaluation cohort.

Expert re-adjudication of the 47 cases initially labelled as 31C revealed substantial annotation inaccuracies. Only 6 cases (12.8%) met the strict criteria for AO/OTA 31C (femoral head fractures). The remaining 41 cases (87.2%) were reclassified as follows: 28 (59.6%) as acetabular fractures (AO/OTA

type 61), 3 (6.4%) as 31A (intertrochanteric fractures), and the remainder as other fracture types or indeterminate classifications.

The final performance data showed that the RAG-LLM framework achieved an accuracy of 94.29% (95% CI: 89.5–98.2%) in distinguishing 31A from 31B fractures on the 70-case test set, with a weighted F1-score of 94.29%. Zero-shot prompting attained the same accuracy (94.29%, F1 = 94.28%), while few-shot learning unexpectedly dropped to 88.57% (F1 = 88.53%) (Table 1).

Table 1
Classification performance of LLM configurations for 31A vs. 31B discrimination

Configuration	Accuracy	F1 Score	31A Precision	31A Recall	31B Precision	31B Recall
Zero-shot	94.29%	94.28%	0.92	0.97	0.97	0.91
Few-shot	88.57%	88.53%	0.94	0.83	0.85	0.94
RAG-LLM	94.29%	94.29%	0.94	0.94	0.94	0.94

Confusion matrices indicated minimal misclassifications under the RAG-LLM condition: among 35 true 31A cases, 33 were correctly classified (94.3%), with 2 misclassified as 31B (5.7%); among 35 true 31B cases, 33 were correctly classified (94.3%), with 2 misclassified as 31A (5.7%). The overall misclassification rate was 5.7% (4 out of 70 cases).

The zero-shot condition exhibited a similar pattern: 1/35 (2.9%) 31A misclassifications and 3/35 (8.6%) 31B misclassifications. The few-shot condition displayed a different error pattern: 6/35 (17.1%) 31A misclassifications and 2/35 (5.7%) 31B misclassifications, suggesting that the provided examples may have biased the model toward 31B predictions (**Fig. 3**).

Figure 3. Confusion matrices for (A) zero-shot, (B) few-shot, and (C) RAG-LLM configurations.

[Insert Fig. 3 here]

Table 2
Subtype classification performance for
31B1 vs. 31B2 discrimination

Metric	Value
Accuracy	70.59%
F1 Score	70.59%
31B1 Recall	60.0% (6/10)
31B2 Recall	85.7% (6/7)
31B1 → 31B2 Errors	40.0% (4/10)
31B2 → 31B1 Errors	14.3% (1/7)

In an exploratory analysis focusing on the 31B subgroup, we evaluated the RAG-LLM framework's ability to distinguish 31B1 (subcapital/non-displaced) from 31B2 (transcervical) fractures using 17 cases with confirmed subtype labels (10 31B1, 7 31B2). The model achieved an accuracy of 70.59% (F1 = 70.59%). The confusion matrix showed 6/10 (60% recall) correct for 31B1 and 6/7 (85.7% recall) correct for 31B2 (Table 2). The predominant error pattern was misclassification of 31B1 as 31B2 (4/10, 40%), indicating that the model struggled to differentiate subcapital from transcervical fractures when reports did not specify the exact anatomical level.

Discussion

This study demonstrates that a RAG-LLM framework can distinguish AO/OTA 31A (intertrochanteric) from 31B (femoral neck) fractures with high accuracy (94.29%) using real-world radiology reports from the MIMIC-IV database. The comparable performance between zero-shot and RAG-augmented prompting suggests that modern LLMs already possess substantial inherent knowledge of AO/OTA classification criteria, while the unexpected decline in few-shot performance highlights the need for careful example selection in clinical NLP applications.

Notably, the most clinically significant finding of this study is the substantial annotation bias identified within the MIMIC-IV dataset. Expert re-adjudication revealed that 87.2% of cases initially labelled as 31C (femoral head fractures) were actually acetabular fractures (AO/OTA type 61) or misclassified 31A/31B fractures. This finding carries important implications for the use of real-world data in AI research and underscores the critical importance of expert data quality assurance.

Compared with previous studies, our results both align with and extend prior work. Inui et al. demonstrated that ChatGPT could classify intertrochanteric fractures into A1 (stable) and A2 (unstable) according to the revised AO/OTA classification system (14). A recent study by Spitzl et al. evaluated ChatGPT-4o and other LLMs in AO fracture classification based on CT text reports, finding that ChatGPT-4o achieved the highest overall accuracy of 74.6% (15). However, these studies were limited by

the use of fabricated reports or small sample sizes, whereas our study extends prior work by using real-world data from a large public database to systematically evaluate the contribution of RAG augmentation(27).

Our finding that RAG did not significantly outperform zero-shot prompting in the primary classification task is consistent with some previous observations that LLMs may already possess sufficient knowledge for common classification tasks (21). Nevertheless, RAG has been shown to substantially improve performance in more complex or uncommon tasks (28). The subtype classification results (70.59%) suggest that RAG may offer greater value in fine-grained discrimination tasks where knowledge completeness is poorer.

Undoubtedly, the high accuracy of the RAG-LLM framework in distinguishing 31A from 31B fractures suggests potential utility in several clinical scenarios. First, automated classification could streamline research workflows by enabling rapid, large-scale phenotyping from electronic health records. Second, in emergency settings, LLM-assisted classification could provide decision support for junior physicians or in resource-limited environments where immediate specialist consultation is unavailable. Third, the framework could be integrated into clinical decision support systems to alert clinicians when radiology report language is ambiguous or inconsistent with expected classification patterns.

Several factors must be considered before clinical deployment: 1) The model was developed and validated on a single-institution dataset and may not generalise to other clinical settings with different reporting conventions(29). 2) The model's performance on more complex or atypical fractures has not been evaluated. 3) Human oversight remains essential, particularly for cases where the model expresses low confidence or where classification would significantly impact treatment decisions (15).

The substantial 31C annotation bias identified in MIMIC-IV carries important implications beyond this study: although MIMIC-IV is one of the most widely used public databases for clinical AI research, researchers should not assume diagnostic labels without independent verification, given that inter-observer agreement is suboptimal—this is particularly critical for studies using AO/OTA or other complex classification systems (30, 31). The high mislabelling rate for 31C (87.2%) may reflect inherent difficulties in distinguishing femoral head from acetabular fractures in radiology reports. Future studies using MIMIC-IV or similar databases should consider implementing multi-stage annotation processes with expert adjudication.

This study has several limitations: 1) The sample size, while adequate for exploratory analysis, was moderate (157 cases, 70 in the test set). Larger, multicentre studies are needed to confirm the generalisability of our findings and to enable more robust subgroup analyses. 2) The study used a single LLM (DeepSeek-Chat); performance may vary with other models or commercial alternatives. 3) The RAG implementation used a static knowledge base rather than a dynamic retrieval system, suggesting that more sophisticated RAG architectures might yield different results. 4) This study focused only on binary classification of 31A versus 31B fractures; performance across the full AO/OTA classification hierarchy, including 31C and subtypes, remains to be evaluated. 5) The subtype analysis was limited by the number

of cases with confirmed subtype labels (n = 17); larger, prospectively annotated datasets are required to assess LLM performance in AO/OTA subtyping. 6) The study did not include comparisons with traditional machine learning methods.

In future research, extending the analysis to include all AO/OTA type 31 fractures (including 31C) and conducting a more comprehensive evaluation of LLM capabilities would be of greatest clinical relevance. Developing multimodal models that integrate both radiology text and imaging data may enhance the model's ability to classify and recognise more complex fractures (32). Prospective studies to further evaluate LLM-assisted workflows could additionally improve temporal and population generalisability.

Conclusion

This study demonstrates that the RAG-LLM framework can distinguish AO/OTA 31A from 31B proximal femoral fractures with excellent performance (94.29% accuracy) using real-world radiology reports from the MIMIC-IV database. The comparable performance between zero-shot and RAG-augmented prompting suggests that modern LLMs already possess substantial inherent knowledge of AO/OTA classification.

The significant 31C annotation bias identified in MIMIC-IV (87.2% mislabelling rate) represents an important contribution to the field, highlighting the critical importance of expert data quality assurance in real-world evidence studies. This finding extends beyond orthopaedics, emphasising the necessity of rigorous annotation protocols in all clinical AI research utilising public databases.

Future work should focus on validation in larger, multicentre datasets, development of multimodal approaches integrating text and imaging data, and assessment of the clinical impact of LLM-assisted fracture classification workflows. With continued refinement and validation, LLM-based systems hold the potential to streamline fracture classification, support clinical decision-making, and facilitate large-scale orthopaedic research using real-world data.

Declarations

Human Ethics and Consent to Participate declarations:

This study utilized data from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database. Access to MIMIC-IV was granted after completion of the required training and credentialing (Certification ID: 77653262). The MIMIC-IV database has been approved by the Institutional Review Boards of the Massachusetts Institute of Technology (MIT) and the Beth Israel Deaconess Medical Center (BIDMC). The requirement for informed consent was waived by both IRBs because the data are de-identified and the research involves no more than minimal risk to participants. This study was conducted in accordance with the Declaration of Helsinki.

Consent to participate:

Not applicable. This study is a retrospective analysis of de-identified data from the MIMIC-IV database. The MIMIC-IV database was approved by the Institutional Review Boards of the Massachusetts Institute of Technology and Beth Israel Deaconess Medical Center, and the requirement for informed consent was waived because the data are de-identified and the research involves no more than minimal risk to participants.

IRB Name

The MIMIC-IV database has been approved by the Institutional Review Boards of the Massachusetts Institute of Technology (MIT) and the Beth Israel Deaconess Medical Center (BIDMC).

Consent for publication

Not applicable.

Availability of data and materials

The data that support the findings of this study are available from the MIMIC-IV database (<https://mimic.mit.edu/>), which is publicly accessible following completion of required data use agreements. Code for the MCP-based multi-agent framework is available from the corresponding author upon reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Shandong Provincial Natural Science Foundation, China (Grant No. ZR2025LMB024). The funder had no role in the study design, data collection, analysis, interpretation of data, writing of the manuscript, or the decision to submit the manuscript for publication.

Authors' contributions

Yuhang Jin: Conceptualization, Methodology, Formal Analysis, Writing.

Zhipeng Peng: Data Curation, Software, Validation, Writing.

Xuesong Yan: Investigation, Resources, Writing.

Jiawei Li: Methodology, Validation, Writing.

Zhuohang Wu: Data Curation, Formal Analysis, Writing.

Yongxian Zhang: Supervision, Project Administration, Writing.

All authors read and approved the final manuscript.

Acknowledgements

The authors acknowledge the Massachusetts Institute of Technology Laboratory for Computational Physiology and the Beth Israel Deaconess Medical Center for providing the MIMIC-IV database.

References

1. Viganò M, Pennestrì F, Listorti E, Banfi G. Proximal hip fractures in 71,920 elderly patients: incidence, epidemiology, mortality and costs from a retrospective observational study. *BMC Public Health*. 2023;23(1):1963.
2. Vazquez O, Belaieff W, Gamulin A, DiGiovanni PL, Cochard B, Hannouche D. Best practices in the management of proximal femoral fractures in elderly patients on the ward: a narrative review. *EFORT open reviews*. 2026;11(3):233-42.
3. Meinberg EG, Agel J, Roberts CS, Karam MD, Kellam JF. Fracture and Dislocation Classification Compendium-2018. *Journal of orthopaedic trauma*. 2018;32 Suppl 1:S1-s170.
4. Song H, Hu S-J, Du S-C, Xiong W-F, Chang S-M. Sub-Classification of AO/OTA-2018 Pertrochanteric Fractures Is Associated With Clinical Outcomes After Fixation of Intramedullary Nails. *Geriatric Orthopaedic Surgery & Rehabilitation*. 2021;12:21514593211056739.
5. Crijns TJ, Janssen SJ, Davis JT, Ring D, Sanchez HB. Reliability of the classification of proximal femur fractures: Does clinical experience matter? *Injury*. 2018;49(4):819-23.
6. Chan G, Hughes K, Barakat A, Edres K, da Assuncao R, Page P, et al. Inter- and intra-observer reliability of the new AO/OTA classification of proximal femur fractures. *Injury*. 2021;52(6):1434-7.
7. Javed J, Islam S. Artificial intelligence and electronic health records: a narrative review of current applications and challenges in pediatric surgery. *World journal of pediatric surgery*. 2025;8(5):e001100.
8. Hossain E, Rana R, Higgins N, Soar J, Barua PD, Pisani AR, et al. Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review. *Computers in Biology and Medicine*. 2023;155:106649.
9. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*. 2023;10(1):1.
10. Ghnemat R, Saleh A. Large language models for clinical artificial intelligence in healthcare a systematic review. *Discover Artificial Intelligence*. 2026;6(1):156.
11. Amin SU, Guizani M, Hossain MS. Advances, Evaluation, and Explainability of Large Language Models in Healthcare: A Systematic Review. 2026;22(2 %J ACM Trans. Multimedia Comput. Commun. Appl.):Article 60.
12. Woźnicki P, Laqua C, Fiku I, Hekalo A, Truhn D, Engelhardt S, et al. Automatic structuring of radiology reports with on-premise open-source large language models. *European Radiology*. 2025;35(4):2018-

29.

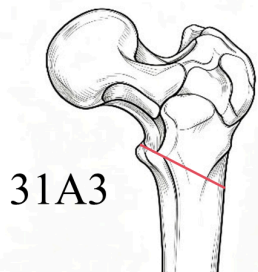
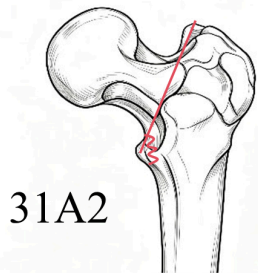
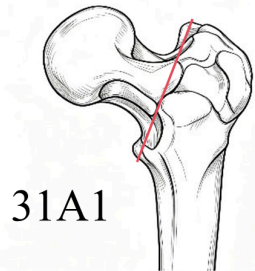
13. Di Palma L, Darvizeh F, Ali M, Fazzini D. Structured Transformation of Unstructured Prostate MRI Reports Using Large Language Models. *Tomography (Ann Arbor, Mich)*. 2025;11(6).
14. Noda M, Takahara S, Hayashi S, Inui A, Oe K, Matsushita T. Evaluating ChatGPT's Performance in Classifying Pertrochanteric Fractures Based on Arbeitsgemeinschaft für Osteosynthesefragen/Orthopedic Trauma Association (AO/OTA) Standards. *Cureus*. 2025;17(1):e78068.
15. Mergen M, Spitzl D, Ketzer C, Strenzke M, Marka A, Makowski M, et al. Leveraging Large Language Models for Accurate AO Fracture Classification from CT Text Reports. *Journal of Imaging Informatics in Medicine*. 2025;39:1861-7.
16. Huang SJIWC. Hallucinations of Large Language Models in Medical Environments: A Systematic Review of Risks, Detection, and Mitigation. 2026;84:03005.
17. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS digital health*. 2025;4(6):e0000877.
18. Hasanzadeh SA, Ghasemian F, Shabaninia E, editors. Bridging Knowledge and Language Models in Healthcare: A RAG Survey. 2025 15th International Conference on Computer and Knowledge Engineering (ICCKE); 2025 28-29 Oct. 2025.
19. Neha F, Bhati D, Shukla DK. Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review. 2025;6(9):226.
20. Zhang G, Xu Z, Jin Q, Chen F, Fang Y, Liu Y, et al. Leveraging long context in retrieval augmented language models for medical question answering. *npj Digital Medicine*. 2025;8(1):239.
21. Wind S, Sopa J, Truhn D, Lotfinia M, Nguyen T-T, Bressem KK, et al. Multi-step retrieval and reasoning improves radiology question answering with large language models. 2025;8.
22. Tayebi Arasteh S, Lotfinia M, Bressem K, Siepmann R, Adams L, Ferber D, et al. RadioRAG: Online Retrieval-Augmented Generation for Radiology Question Answering. 2025;7(4):e240476.
23. Mao Y, Wang C, Li Y, Wang W, Zhang M. Multidimensional evaluation of large language models in radiology report readability. *npj Digital Medicine*. 2026;9(1):585.
24. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*. 2023;10(1):1.
25. Xu P, Wang L. Evaluating the impact of Age-Adjusted Charlson Comorbidity Index on in-hospital complications in patients with femoral fracture: a retrospective cohort analysis from the MIMIC-IV 2.2 database. 2025;Volume 12 - 2025.
26. DeepSeek-AI. DeepSeek-V3 Technical Report. arXiv:2412.19437.
27. Hu S, Keeley T, Halvorson R, Campbell ST, Lefaivre KA, Levack A, et al. Deriving the OTA/AO fracture classification from routinely collected radiology reports using a large language model. *OTA international : the open access journal of orthopaedic trauma*. 2026;9(2):e482.

28. Wang Z, Wu D, Nguyen Q, Wang K. Integrating Chain-of-Thought and Retrieval Augmented Generation Enhances Rare Disease Diagnosis from Clinical Notes. ArXiv. 2026.
29. Kim SH, Schramm S, Adams LC, Braren R, Bressemer KK, Keicher M, et al. Benchmarking the diagnostic performance of open source LLMs in 1933 Eurorad case reports. npj Digital Medicine. 2025;8(1):97.
30. Chang T, Sjoding MW, Wiens J. Disparate Censorship & Undertesting: A Source of Label Bias in Clinical Machine Learning. Proceedings of machine learning research. 2022;182:343-90.
31. Ramadan B, Liu M-C, Burkhart MC, Parker WF, Beaulieu-Jones BK. Diagnostic Codes in AI Prediction Models and Label Leakage of Same-Admission Clinical Outcomes. JAMA Network Open. 2025;8(12):e2550454-e.
32. Bouslimi R. A knowledge-guided and uncertainty-calibrated multimodal framework for fracture diagnosis and radiology report generation. Computerized Medical Imaging and Graphics. 2026;128:102709.

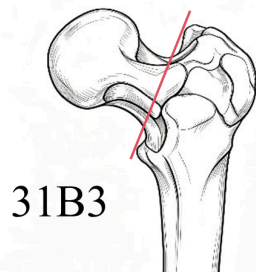
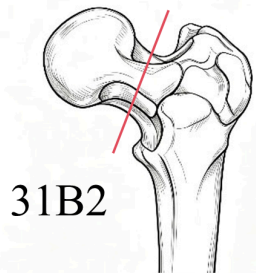
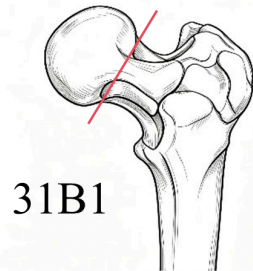
Figures

AO/OTA 31-Type Proximal Femoral Fracture Classification

31A Intertrochanteric



31B Femoral Neck



31C Femoral Head

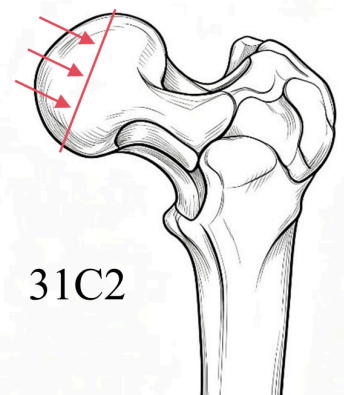
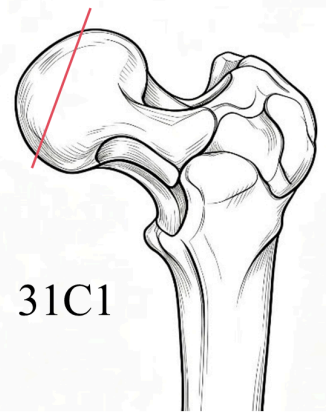


Figure 1

AO/OTA type 31 classification of proximal femoral fractures.

Confusion Matrices: Classification Performance Comparison

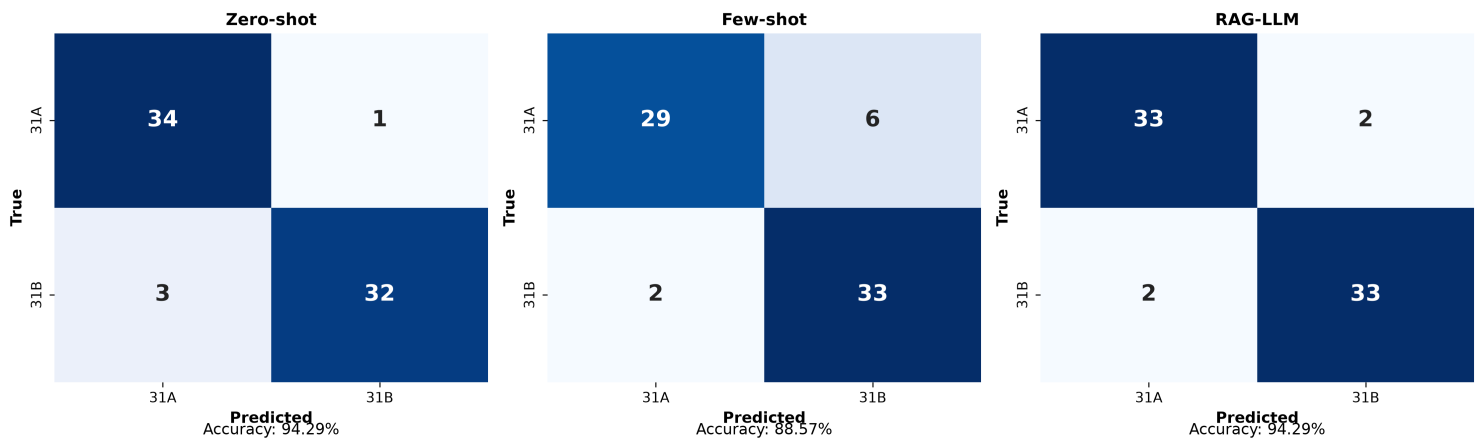


Figure 2

Flowchart of case identification, annotation, and dataset preparation.

Study Flow Diagram: Case Identification and Dataset Preparation

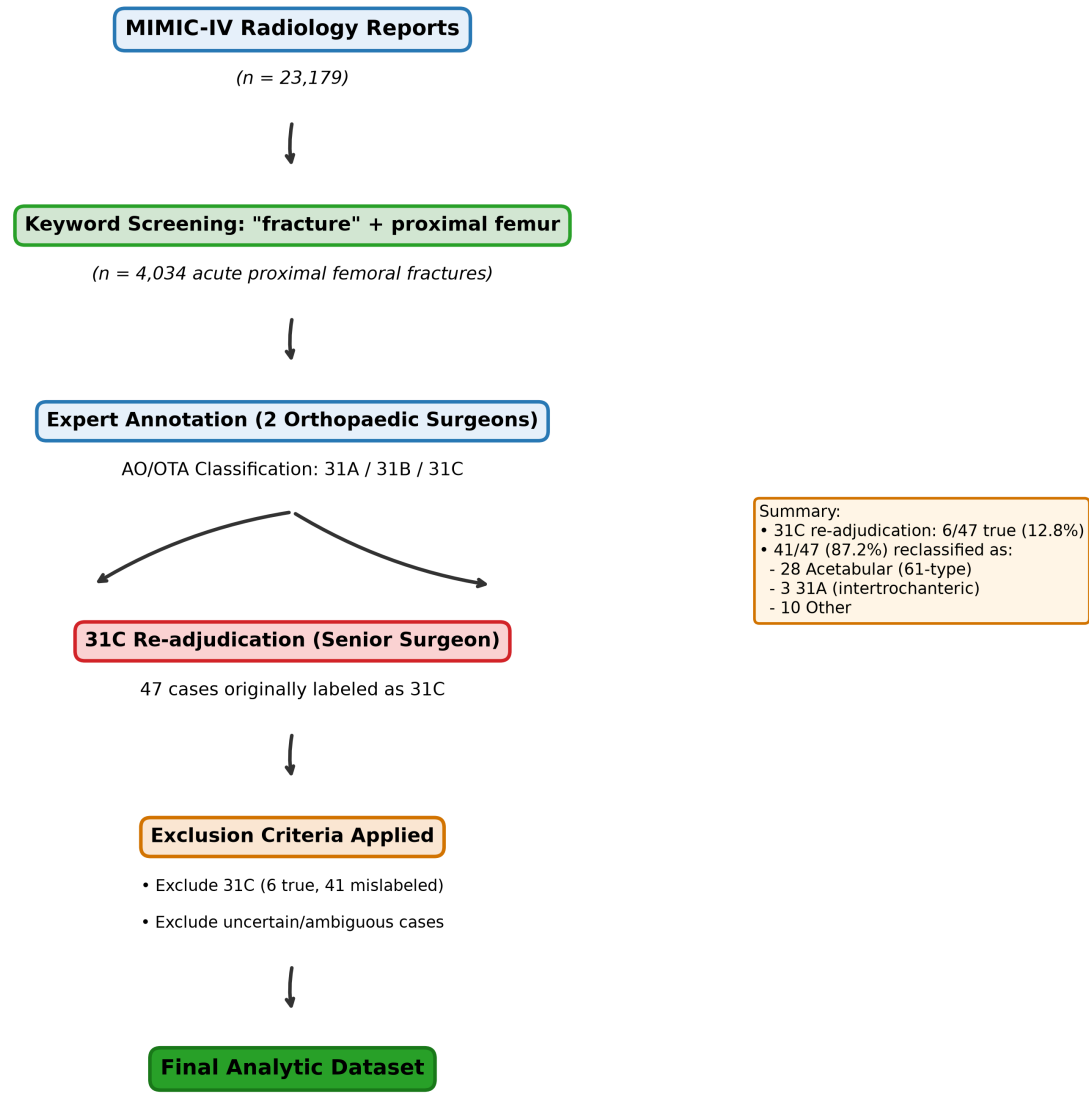


Figure 3

Confusion matrices for (A) zero-shot, (B) few-shot, and (C) RAG-LLM configurations.