HealthManagement
Promoting Management and Leadership

Automated LLM Labelling Drives Multi-Label Radiography

Automating the conversion of routine radiology reports into structured training data remains a bottleneck for imaging AI. An evaluation of a large language model (LLM) applied to upper extremity radiography examined whether zero-shot label extraction from free-text reports could provide accurate, uncertainty-aware annotations for multi-label image classification. Radiography series of the clavicle (n=1170), elbow (n=3755) and thumb (n=1978) were processed after anonymisation, with labels assigned as present, absent or uncertain. Extracted labels then trained convolutional neural networks (CNNs) for each anatomical region. Accuracy was verified on internal and external test sets, and the influence of handling uncertainty during training was assessed. The approach sought to accelerate dataset creation while maintaining diagnostic performance and generalisability across sites.

Two-Centre Pipeline from Reports to Training Labels

A retrospective, two-centre design combined radiography and corresponding reports from a university hospital for internal training, validation and testing and an external hospital for independent testing. Reports were in German and anonymised before processing. OpenAI's GPT-4o operated in a zero-shot mode to complete predefined JavaScript Object Notation templates for each region, designed by a senior musculoskeletal radiologist to capture frequent and less common conditions. For every condition, the LLM selected true, false or uncertain based on report wording.

Manual verification was performed for all internal and external test sets. Internal test cohorts comprised 233 clavicle reports, 745 elbow reports and 393 thumb reports, external cohorts comprised 300 reports per region. In the test sets, all labels were finalised as true or false, with follow-up imaging used where needed. Across the combined test sets, automatic extraction was correct in 98.6% of labels (60,618 of 61,488). Region-specific label-level accuracies for the external cohorts reached 98.6% for clavicle, 98.4% for elbow and 98.1% for thumb, with report-level accuracies between 71.3% and 73.7%. Internal cohorts showed similar label-level accuracies of 98.6% to 99.0%, with report-level accuracies between 74.4% and 85.5%.

Label uncertainty was present but relatively infrequent. In internal cohorts, manual review identified uncertain wording in 3.9% of clavicle reports, 10.5% of elbow reports and 9.7% of thumb reports, while the extraction pipeline automatically flagged 0.9%, 6.4% and 5.3%, respectively. External cohorts showed a similar pattern, with uncertainty present in 5.3% of clavicle, 16.3% of elbow and 16.0% of thumb reports, of which 3.3%, 8.7% and 13.3% were detected automatically.

Handling Uncertainty with Inclusive and Exclusive Strategies

To test the operational impact of uncertainty during model development, two training strategies were compared while keeping all other parameters constant. In the inclusive approach, labels originally marked as uncertain in training and validation were reassigned to true. In the exclusive approach, uncertain labels were reassigned to false. All test sets contained only definitive labels to allow unbiased evaluation. The volume of uncertain labels in training and validation was limited (42 for clavicle, 492 for elbow and 231 for thumb).

Models were implemented in PyTorch using a modified ResNet50 backbone configured for multi-label output with sigmoid activation. For elbow and thumb, anteroposterior and lateral projections were processed by separate networks with feature concatenation before classification. Standard augmentation and resizing to 512×512 pixels were applied. Operating thresholds on the test sets were chosen using the Youden index derived from validation performance.

Must Read: ***Out-of-the-Box LLMs Flag Critical Radiology Findings***

Comparisons used macro-averaged receiver operating characteristic area under the curve (AUC), calculated across labels with at least 10 positive cases per test set. Statistical testing employed the DeLong method with Benjamini–Hochberg correction. Across regions and datasets, no significant AUC differences were observed between inclusive and exclusive training or between internal and external testing ($p \geq 0.15$). This indicated that the chosen treatment of uncertain labels during training did not materially alter downstream diagnostic performance in this setting.

Performance and Generalisation Across Clavicle, Elbow and Thumb

Performance varied by region and label prevalence but remained competitive overall. For clavicle, macro-averaged AUC reached 0.80 (range 0.59–0.95) for inclusive training and 0.81 (0.63–0.94) for exclusive training. For elbow, macro-averaged AUC was 0.80 (0.62–0.87) for inclusive and 0.80 (0.61–0.88) for exclusive. For thumb, macro-averaged AUC was 0.76 (0.59–0.91) for inclusive and 0.78 (0.61–0.90) for exclusive. External generalisation was maintained, exemplified by elbow models reaching 0.79 macro-averaged AUC for both strategies.

Label-level behaviour reflected class balance and visual conspicuity. Fracture-related labels and displacement exhibited high AUCs where positive case counts were larger. Rarer or subtler findings, particularly soft-tissue abnormalities and small ossicles, showed lower or less consistent AUCs and wider confidence intervals. Threshold-dependent metrics like sensitivity and specificity varied by label, with some trade-offs evident at the single Youden-selected operating point, yet threshold-independent AUC comparisons remained stable across strategies and sites.

The uncertainty detection gap between manual review and automated extraction did not translate into measurable AUC differences. Given the low absolute proportion of uncertain labels in training and validation, the effect size of reassigning them as positive versus negative appears limited under the tested conditions. Together with the high label-level extraction accuracy, these results support the feasibility of zero-shot LLM labelling for assembling multi-label training datasets from routine reports in upper extremity radiography.

Zero-shot LLM extraction produced high-accuracy structured labels from routine radiology reports and enabled competitive multi-label CNNs for clavicle, elbow and thumb radiography. Performance generalised to an external site, and alternative strategies for handling uncertain wording during training did not significantly affect results. Stronger performance for common fracture-related findings and weaker performance for rarer soft-tissue labels underline the continued importance of class balance and case availability. The workflow provides a scalable way to create training labels from routine reports, shortening data preparation cycles for imaging AI and retaining clinically relevant performance.

Source: **European Radiology**

Image Credit: **iStock**

Published on : Sun, 23 Nov 2025