SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Evaluating the Diagnostic Value of a Large Language Model in Analyzing Chest X-Ray and CT Imaging in ED Trauma Cases

David Choung, University of California, Irvine; Catherine Kim, University of California, Irvine; Shawn Sun, MD; Dillon Sommer; Roozbeh Houshyar, MD

Introduction/Background
Overutilization of imaging methods – particularly chest x-rays (CXR) and computed tomography (CT) -- has become an important consideration in the care of trauma patients in the emergency department (ED) setting. Utilizing a large language model (LLM), this retrospective study examines trauma cases to assess CXR's diagnostic value by analyzing the frequency of acute findings that led to acute changes in clinical management. Furthermore, this study explores the potential of an LLM in automating the annotation process.

Methods/Intervention
A total of n = 155 patient cases were reviewed from a singular institution's ED in southern California from January to December of 2024. Two human annotators independently evaluated each case for: (1) presence of acute findings (i.e. tension pneumothorax, hemothorax, tube placement (endotracheal tubes, chest tubes, central lines, PermaCath), pericardial/cardiac tamponade, pericardial effusion, penetrating injury, and/or airway obstruction), (2) relevant interventions performed between CXR and CT, and (3) whether imaging contributed added clinical value to patient management. An OpenAI LLM was prompted with the same 155 cases, and the outputs were compared to determine accuracy in identifying acute findings, interventions, and clinical value added.

Results/Outcome
Compared to human annotations, LLM annotations achieved 98.06% (152/155), 91.61% (142/155), and 88.39% (137/155) accuracy across acute findings, interventions, and value added respectively. McNemar's test showed no significant difference between human and LLM annotations for both acute findings and interventions (p-value = 0.5637) but showed significant difference in assessing added clinical value (p-value = 0.0003).

Conclusion
Human-generated, manual annotations and LLM annotations were in concordance across the detection of acute findings and pertinent interventions between CXR and CT. Conversely, there existed a statistically significant difference in determining the added diagnostic value and clinical impact. LLM's can be highly effective in managing clinical decision making and reducing workload burden in a critical care setting yet potentially limited in grasping nuanced clinical judgement to alter patient management.

Statement of Impact
Our study underscores the potential benefit of integrating LLM into radiology for accurate, rapid detection of acute findings and guidance of clinical interventions, with potential for improved accuracy in decision making.

	LLM = YES	LLM = NO
Human = YES	67	2
Human = NO	1	85

Table 1: Human vs. LLM Agreement in Annotation of Acute Findings

Table 1: Human vs. LLM Agreement in Annotation of Acute Findings

	LLM = YES	LLM = NO
Human = YES	5	5
Human = NO	7	138

Table 2: Human vs. LLM Agreement in Annotation of Interventions

Table 2: Human vs. LLM Agreement in Annotation of Interventions

	LLM = YES	LLM = NO
Human = YES	54	1
Human = NO	16	84

Table 3: Human vs. LLM Agreement in Annotation of Value Added

Table 3: Human vs. LLM Agreement in Annotation of Value Added

Keywords
Radiology; LLM (large language model); Trauma; CXR; CT; Clinical decision making