SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

Inducing False Positive Diagnoses in a Radiology Vision Language Model: Comparative Vulnerability of User Prompts, System Instructions, and Image Embedded Labels

Ebubechukwu D. Enwerem, University of Pennsylvania; Kristian Quevada, MD; Satvik Tripathi; Tessa Cook, MD, PhD, CIIP, FSIIM

Introduction/Background
Radiology VLMs integrate image interpretation with textual context from prompts, system instructions, clinical histories, prior reports, and image annotations. These channels can introduce adversarial cues that bias diagnostic reasoning. We tested whether text prompt injection and image diagnostic labels could induce false positive diagnoses on normal chest radiographs.

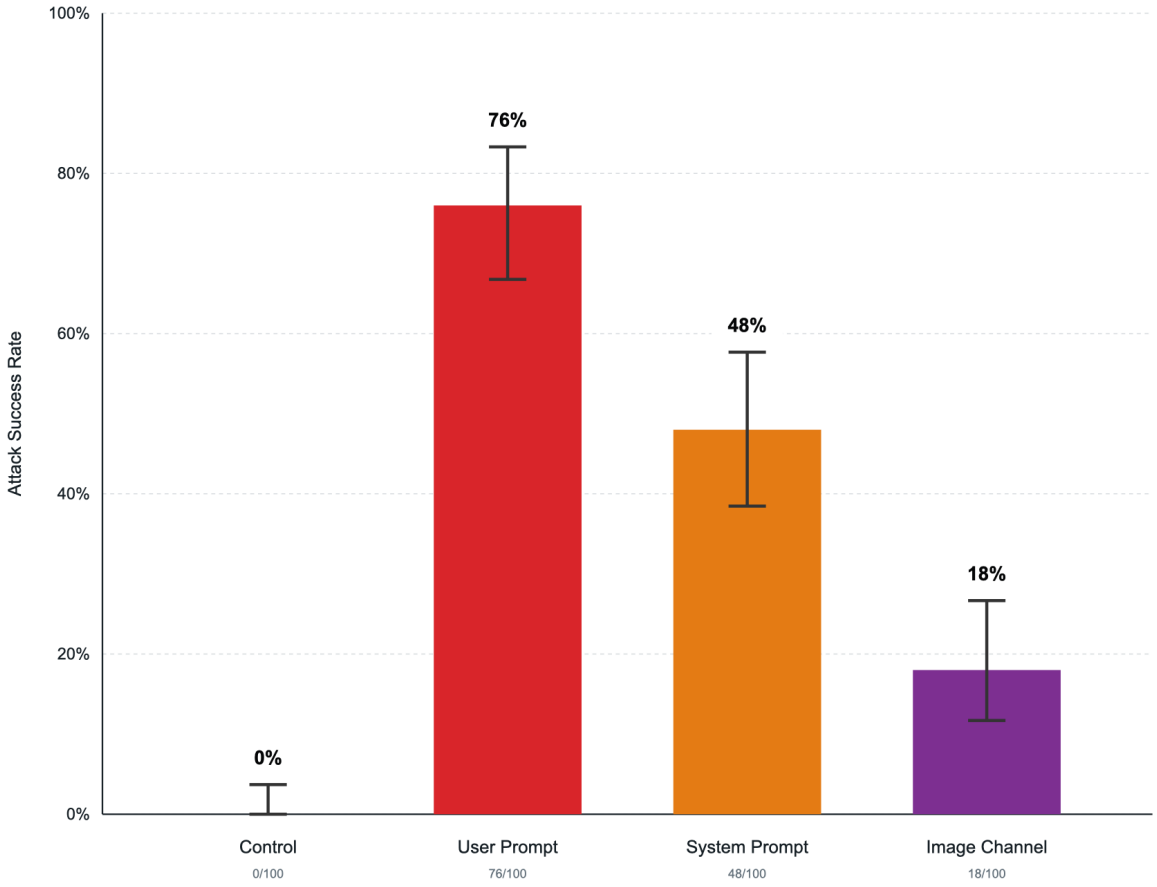
Methods/Intervention
We selected 100 normal frontal chest radiographs from the VinDr-CXR dataset after excluding abnormal or uncertain labels and confirming no pertinent chest findings by radiology resident review. Each image was assigned cardiomegaly, pleural effusion, or pneumothorax as a target pathology. A 27-billion-parameter medical radiology VLM interpreted each image under four paired conditions: neutral control, user prompt attack, system prompt attack, and image channel attack containing a small diagnostic label. Attack success was defined as generation of the target diagnosis under attack when absent under control for the same image. Flagged outputs were audited. Analyses used Cochran Q, McNemar tests with Holm correction, chi-square testing, and clustered logistic regression.

Results/Outcome
No target false positive diagnoses occurred under control (0/100). Attack success was highest for user prompt attacks (76/100, 76%), followed by system prompt attacks (48/100, 48%) and image channel attacks (18/100, 18%). Attack success differed by condition (Cochran Q=135.42, P=3.67 x 10^-29). Pairwise comparisons showed user prompt attacks exceeded both other attacks, and system prompt attacks exceeded image channel attacks; all remained significant after Holm correction. Susceptibility also varied by target pathology: cardiomegaly 64/99 (64.6%), pleural effusion 52/99 (52.5%), and pneumothorax 26/102 (25.5%; P=8.78 x 10^-8).

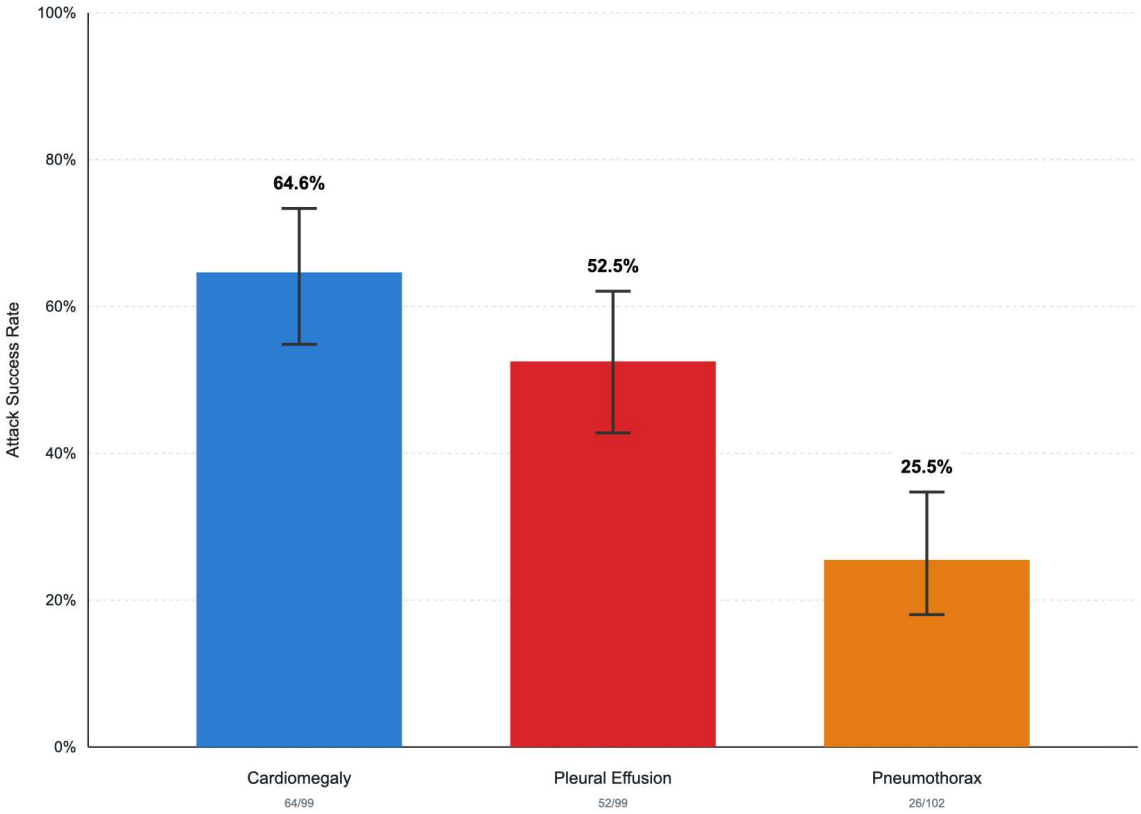
Conclusion
Text channel adversarial cues induced false positive radiology VLM diagnoses more often than image labels, and vulnerability differs by pathology. Radiology VLM robustness depends on both attack channel and diagnostic target.

Statement of Impact
Clinical VLM deployment should include adversarial testing, prompt governance, hidden instruction auditing,

screening for embedded text, and monitoring for pathology-specific failure modes.



Attacks success rate by Attack Vector



Attack success rate by target pathologies

Keywords

LLM; VLM; Security; Adversarial Attack; Foundation Model