

Supplementary Table 1. Performance of voting ensemble from all 3 models using Simple prompting.

Thresholds (No. of votes)	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)	F1-score (%)
Voting ensemble for Acute PE					
≥1	97.0	96.7	98.2	96.9	97.6
≥2	93.9	98.4	99.0	95.5	96.4
≥3	90.6	98.9	99.3	93.6	94.8
Voting ensemble for Bilateral PE					
≥1	99.3	65.2	71.5	81.2	83.1
≥2	98.7	85.4	85.5	91.6	91.6
≥3	97.0	90.4	89.8	93.5	93.3
Voting ensemble for Large PE					
≥1	91.4	60.0	47.5	68.9	62.5
≥2	81.3	81.0	62.8	81.1	70.9
≥3	77.5	86.5	69.4	83.9	73.2

Supplementary Table 2. Performance of voting ensemble from all 3 models using Chain-of-Thought prompting.

Thresholds (No. of votes)	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)	F1-score (%)
Voting ensemble for Acute PE					
≥1	94.4	74.2	86.9	87.2	90.5
≥2	92.0	98.1	98.9	94.1	95.3
≥3	87.9	98.9	99.3	91.8	93.2
Voting ensemble for Bilateral PE					
≥1	99.3	68.4	73.4	82.9	84.4
≥2	97.7	90.1	89.6	93.6	94.5
≥3	96.7	92.1	91.5	94.2	94.0
Voting ensemble for Large PE					
≥1	88.2	66.6	51.1	72.7	64.7
≥2	81.8	82.0	64.3	82.0	72.0
≥3	75.4	87.3	70.1	83.9	72.7

Natural Language Processing for PE Classification

All reports underwent standardized preprocessing prior to rule application in Python: text was converted to lowercase, null characters were removed, repeated whitespace was collapsed, and leading/trailing spaces were trimmed. When present, the ‘Impression,’ ‘Conclusion,’ ‘Summary,’ or ‘Opinion’ section was preferentially extracted using header-based heuristics; otherwise, the entire report text was analyzed. Pattern matching was implemented using regular expressions with context-aware negation handling. A predefined lexicon of negation cues (e.g., “no,” “not,” “without,” “negative,” “absence of,” “ruled out,” “exclude”) was applied within bounded character windows surrounding candidate matches, and sentence boundaries were approximated using punctuation delimiters. Contrastive cues (e.g., “however,” “but,” “though,” “yet,” “except”) were used to limit the scope of negation to the most relevant clause. For acute PE detection, positive search terms included phrases such as “acute pulmonary embolism,” “acute PE,” “positive for pulmonary embolism,” “consistent with pulmonary embolism,” “evidence of pulmonary embolism,” “filling defect in the pulmonary artery,” “thrombus in the main pulmonary artery,” and anatomic modifiers (e.g., segmental, subsegmental, lobar, main, central) directly linked to embolism terminology. Negative search terms included “no pulmonary embolism,” “negative for pulmonary embolism,” “without evidence of pulmonary embolism,” “no PE,” and “no filling defect.” Chronic expressions (e.g., “chronic PE,” “prior embolism,” “resolved PE”) and uncertainty phrases (e.g., “cannot exclude PE,” “equivocal for

PE,” “possible PE”) were also identified. The decision tree for acute PE classified reports as negative if a negated PE pattern was present without strong non-negated positive evidence; classified as positive if strong positive evidence appeared outside negated context; classified as negative if only chronic language without acute qualifiers was present; classified as negative if only uncertainty language was present; and otherwise defaulted to negative.

For bilateral PE (applied only within PE-positive reports), bilateral involvement was identified if non-negated explicit bilateral terms such as “bilateral pulmonary emboli,” “embolism in both lungs,” or “PE in both pulmonary arteries” were present. In the absence of explicit bilateral wording, bilateral classification required separate non-negated clauses demonstrating right-sided PE anatomy (e.g., “right pulmonary artery embolus,” “right lobar PE”) and left-sided PE anatomy (e.g., “left pulmonary artery embolus,” “left segmental PE”), with laterality explicitly tied to embolism-related anatomic structures (e.g., pulmonary artery, lobar, segmental, subsegmental) rather than unrelated findings. If neither explicit bilateral language nor both right- and left-sided PE evidence were identified, the case was classified as non-bilateral.

For PE size stratification within confirmed PE cases, large PE search terms included “saddle pulmonary embolism,” “main pulmonary artery embolus,” “right main pulmonary artery,” “left main pulmonary artery,” “pulmonary trunk,” “lobar embolus,” whereas small PE terms included “segmental,” “subsegmental,” and explicit burden descriptors such as “small PE” or “tiny embolus.” The size decision logic prioritized proximal anatomy: if both large and small anatomic descriptors were present, the case was classified as large; if large anatomic terms were present alone, it was classified as large; if only segmental or subsegmental terms were present, it was classified as small; if only explicit small-burden descriptors were present, it was classified as small; and if no size descriptors were identified, the default classification was small.