

Automating the Management of Extra-Spinal Findings in MRI Spine Studies using a Privacy-Preserving Large Language Model: A Single-Institution Feasibility Study

Gifford Tan^{1,*}, Wilson Ong^{1,*}, Yonghan Ting^{1,2}, Shuliang Ge^{1,2}, Yi Liang Tan^{1,2}, Xi Zhen Low^{1,2}, Wei Chuan Tan¹, Andrew Makmur^{1,2}, Naomi Wenxin Leow³, Muzamil Arif Din Abdul Jabbar^{4,5}, Qai Ven Yap⁶, Shao Jin Ong^{1,2}, Jiong Hao Jonathan Tan⁷, Naresh Kumar⁷, James Thomas Patrick Decourcy Hallinan^{1,2}

¹ Department of Diagnostic Imaging, National University Hospital, 5 Lower Kent Ridge Rd, Singapore 11907 ² Department of Diagnostic Radiology, Yong Loo Lin School of Medicine, National University of Singapore, 10 Medical Drive, Singapore, 117597 ³ AIO Innovation Office, National University Health System, 3 Research Link, #02-04 Innovation 4.0, Singapore 117602 ⁴ School of Clinical Medicine, University of Cambridge, Hills Rd, Cambridge CB2 0SP. ⁵ Agency for Science, Technology and Research (A*STAR), 2 Fusionopolis Way Singapore 138634. ⁶ Biostatistics Unit, Yong Loo Lin School of Medicine, National University of Singapore, Singapore ⁷ National University Spine Institute, Department of Orthopaedic Surgery, National University Health System, 1E, Lower Kent Ridge Road, 119228, Singapore

Both GT and WO contributed equally to this manuscript and are co-first authors.

Background and Objective

The problem statement

MRI spine studies frequently reveal clinically relevant extra-spinal findings (ESFs) that may require further evaluation or specialist referral.

We currently rely on doctors to manually review radiology reports and electronic medical records (EMRs) to determine whether these ESFs require further management.

Objective

This study aims to evaluate the use of a privacy-preserving large language model (PP-LLM) to automate the management of extra-spinal findings reported in MRI spine studies.



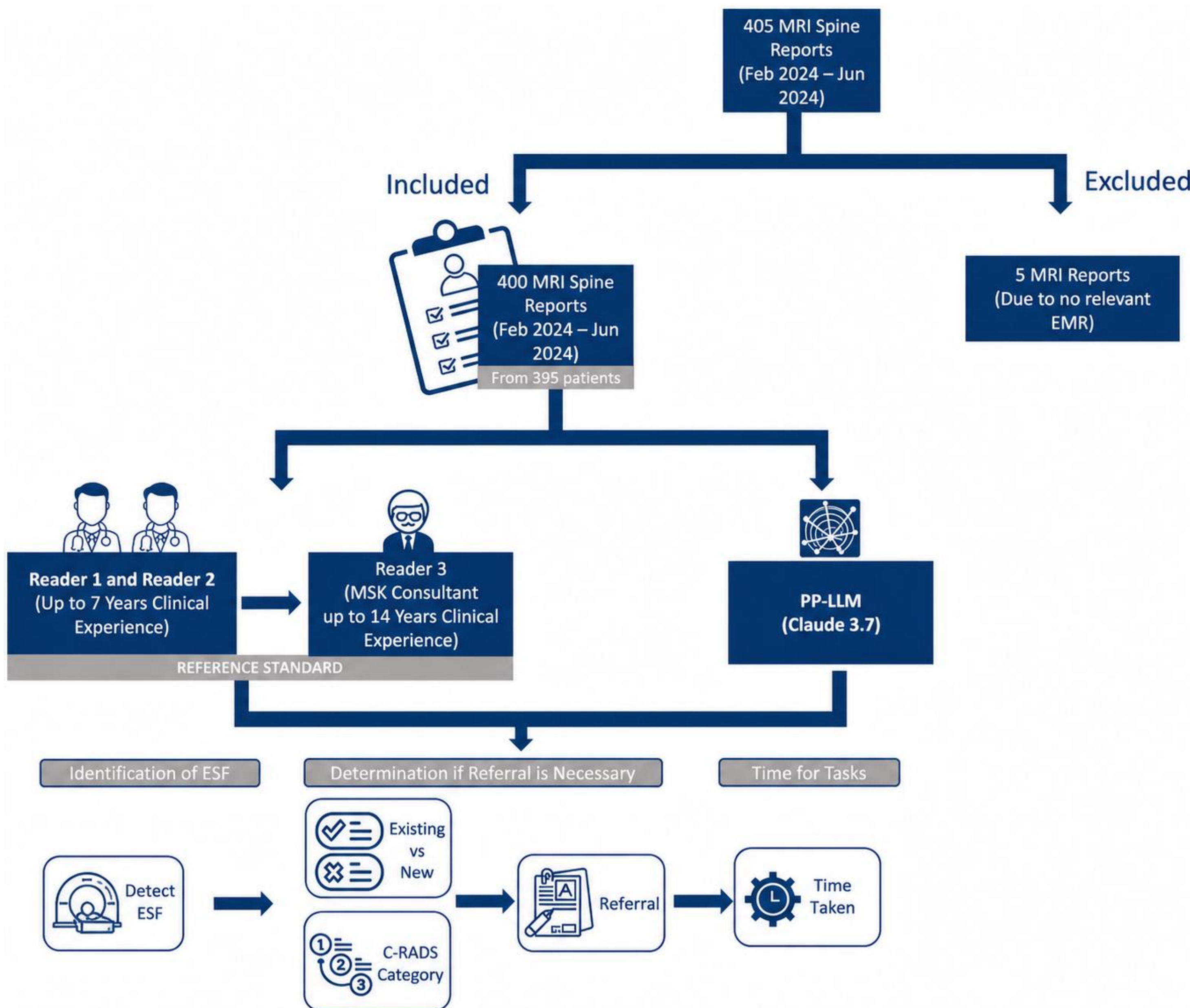
Methods

Establishing ground truth – What are we assessing the PP-LLM on?

400 MRI reports were first reviewed by two clinicians (Reader 1 and Reader 2) to evaluate for the following 4 tasks. Discrepancies were resolved by a third independent reviewer (Reader 3) to establish the reference standard.

Can Reader 1, Reader 2 and the PP-LLM:

- (1) Detect the ESF in the MRI report
- (2) Classify novelty (or if it is pre-existing)
- (3) Triage clinical significance
- (4) Assess referral necessity and generate appropriate referral



Assessing clinical significance with modified C-RADS

The CT Extracolonc Findings Reporting and Data System (C-RADS), was adapted to determine the clinical significance of ESFs.

Modified C-RADS	Definition	Example
E1/E2	No clinically important findings or benign findings that require no additional workup	Non-obstructing renal calculi Extra-renal pelvis
E3	Indeterminate, incompletely characterized findings, for which clinical correlation and further work-up and subspecialty referral could be performed if indicated.	Simple or complicated renal cyst Renal contour abnormality, cannot exclude exophytic mass
E4	Potentially important findings requiring further work-up and early communication to the physician.	Staghorn kidney calculus or other urolithiasis causing obstruction Suspicious renal mass or lesion

Results

PP-LLM detected 99.8% of all 401 ESFs in seconds Outperformed human readers in C-RADS classification

Extraspinal ESF Detected	Reader 1	Reader 2	PP-LLM
Total (n = 401)	401/401 (100.0%)	399/401 (99.5%)	400/401 (99.8%)
ESF Classified			
C-RADS E1/E2 (n = 95, 23.7%)	101 (25.2%)	100 (24.9%)	98 (24.4%)
C-RADS E3 (n = 294, 73.3%)	288 (71.8%)	287 (71.6%)	290 (72.3%)
C-RADS E4 (n = 12, 3.0%)	12 (3.0%)	12 (3.0%)	12 (3.0%)
Overall Accuracy	383/401 (95.5%)	380/401 (94.8%)	388/401 (96.8%)

No statistical difference between PP-LLM and human Readers 1 and 2 Outperformed human readers in identifying if ESF is new

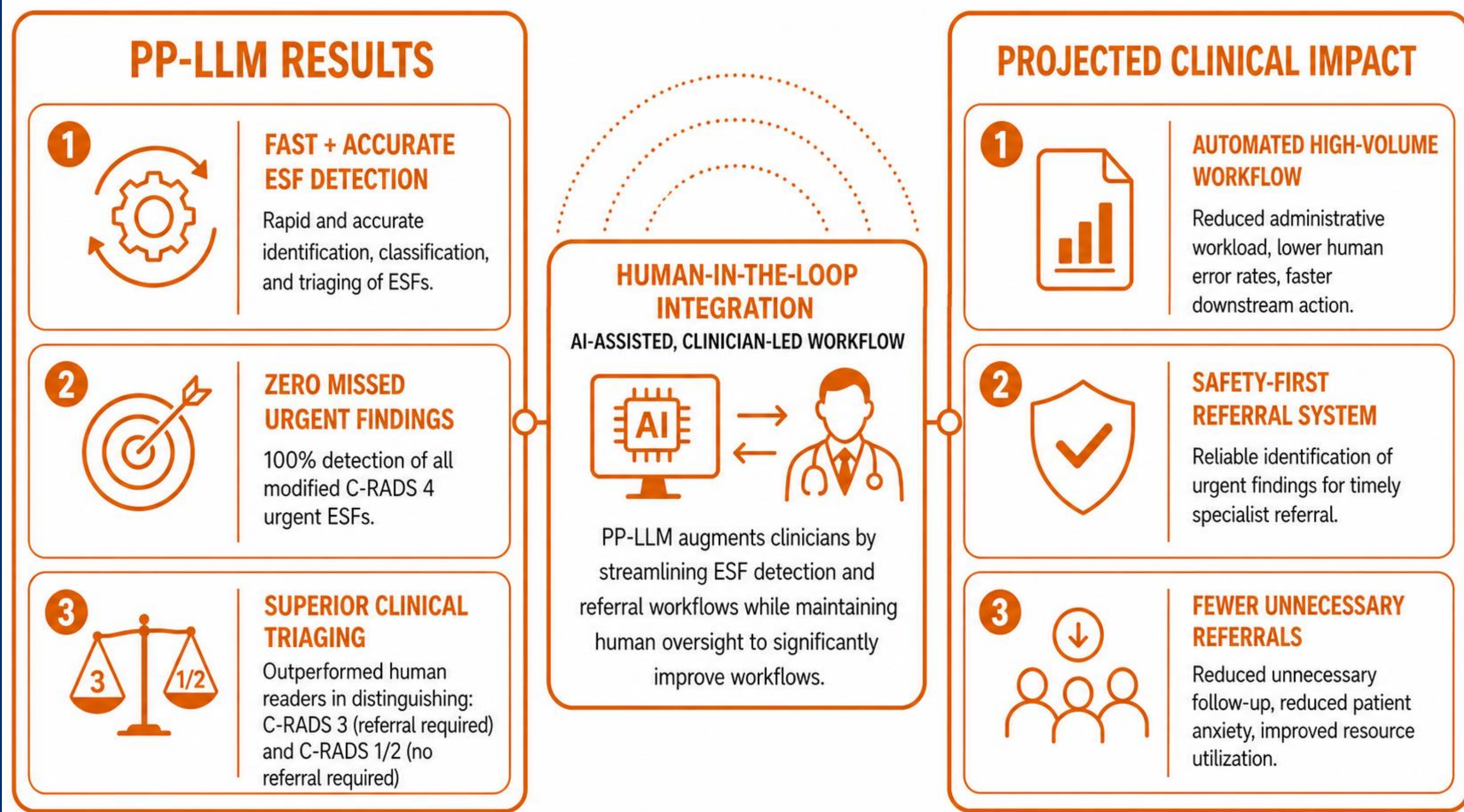
	PP-LLM vs Reader 1	PP-LLM vs Reader 2
Accuracy of C-RADS Classification	96.8 % vs 95.5% (McNemar's p-Value = 0.359)	96.8 % vs 94.7% (McNemar's p-Value = 0.152)
Accuracy of Determining if ESF is new	98.5 % vs 95.3% (McNemar's p-Value = 0.200)	98.5 % vs 94.0% (McNemar's p-Value <0.01)
Specificity of Determining if ESF is new	94.2 % vs 69.4% (McNemar's p-Value <0.01)	94.2 % vs 67.3% (McNemar's p-Value <0.01)

PP-LLM detected all urgent E4 findings with 100% accuracy, sensitivity and precision

PP-LLM Performance on classifying ESF into various C-RADS category.								
Category	True Positive	True Negative	False Positive	False Negative	Accuracy	Sensitivity	Precision	F1 Score
E1/E2	90	298	8	5	96.8%	94.7%	91.8%	93.2%
E3	286	103	4	8	97.0%	97.3%	98.6%	98.0%
E4	12	389	0	0	100.0%	100.0%	100.0%	100.0%

Discussion

Clinical implications and feasibility



Potential Use Case

