

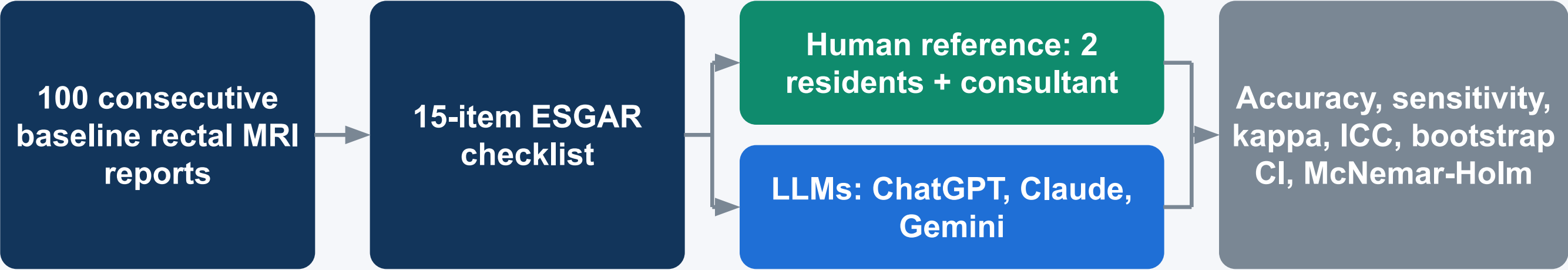
Leveraging Large Language Models for Automated Quality Assessment of Baseline Rectal MRI Reports Using ESGAR Reporting Recommendations

Shahroze Ahmed, Fatima Baluch, Kashif Siddique
Shaukat Khanum Memorial Cancer Hospital & Research Centre, Lahore, Pakistan

Objective

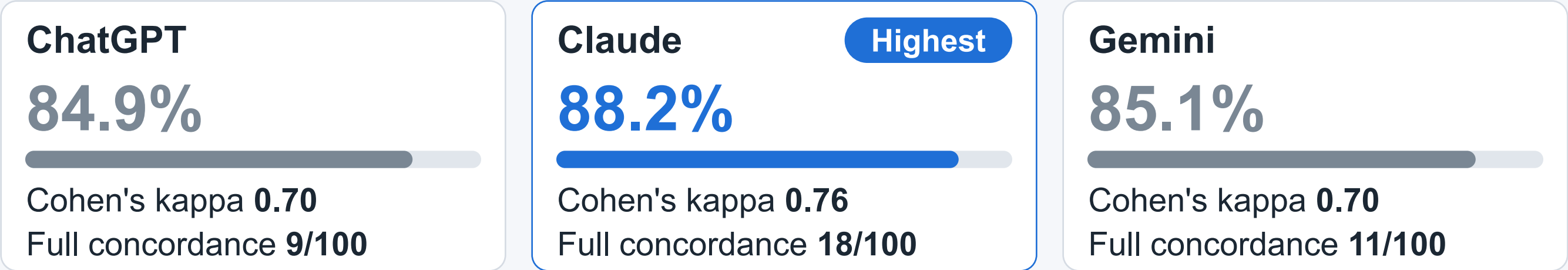
Although ESGAR recommends standardized rectal MRI reporting, routine reports often remain free-text and incomplete, making manual quality auditing time-consuming. This study evaluated the performance of three large language models (ChatGPT, Claude, and Gemini) in assessing the completeness of baseline rectal MRI reports against a human reference standard using an ESGAR-derived reporting checklist.

Methods



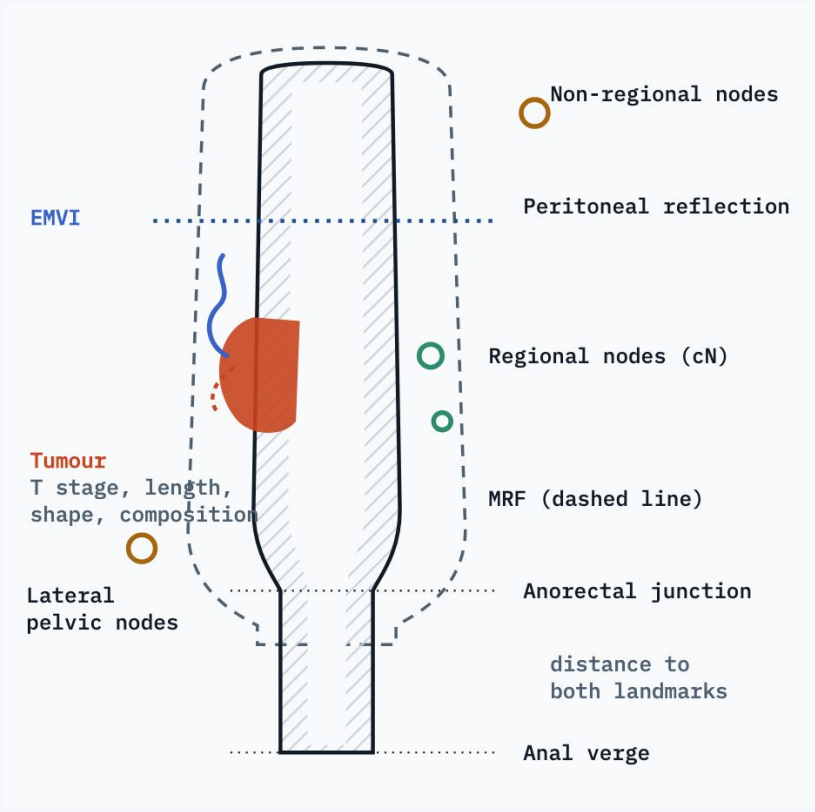
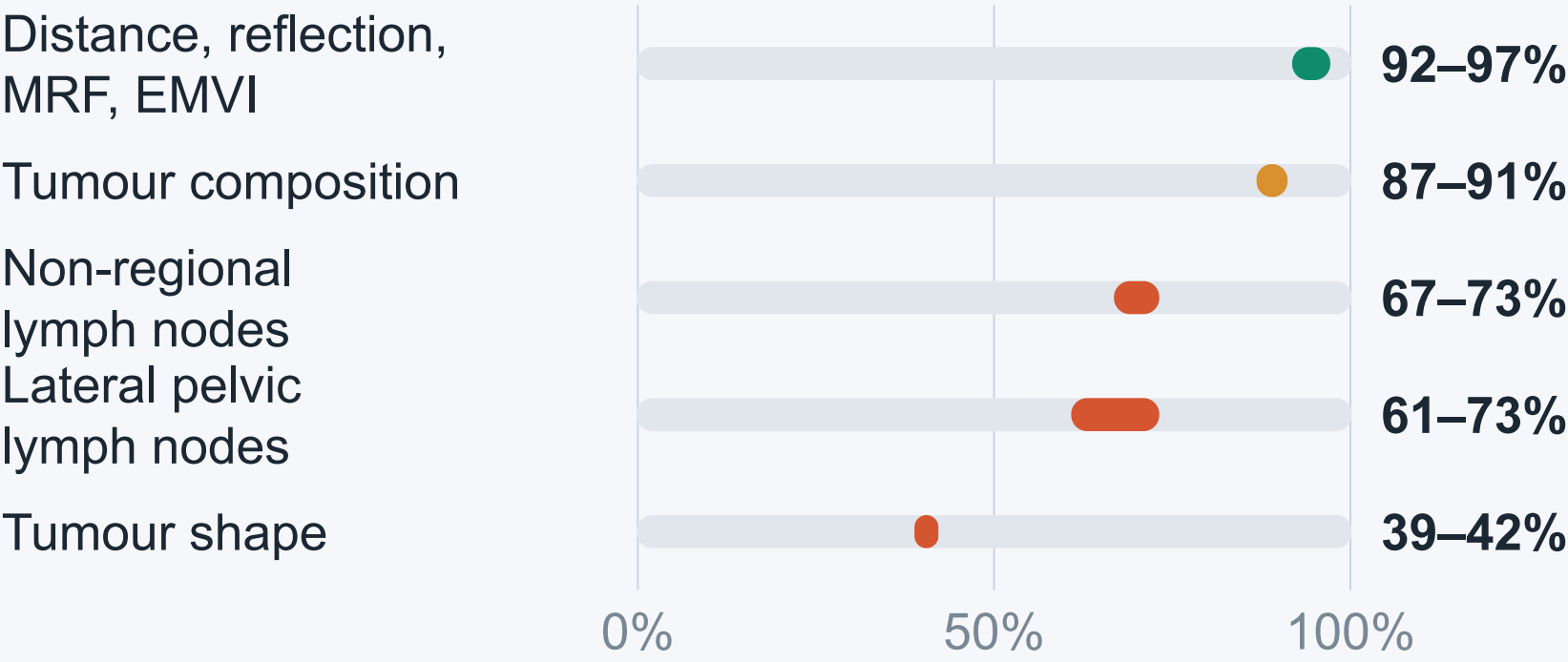
Retrospective; identical documentation-focused prompt for every model. Paired Wilcoxon tests; exact McNemar tests with Holm correction.

Results: overall agreement (1,500 checklist assessments per model)



Claude: highest sensitivity (90.6%) and kappa. ChatGPT: highest specificity (88.3%).

Accuracy by checklist item and key statistics



Conclusion

All three LLMs showed moderate-to-substantial agreement with human assessment (kappa 0.70–0.76), and Claude had the highest accuracy and kappa (88.2% accuracy, κ 0.76). Staging elements were identified reliably, but less frequently documented items were not. LLMs show promise for report quality auditing, pending standardized prompts and multicentre validation.

References

- Beets-Tan RGH, et al. Eur Radiol. 2018;28:1465–1475.
- ESGAR Rectal Imaging Guideline Group. Eur Radiol. 2026;36:4592–4607.
- Gertz RJ, et al. Radiology. 2024;311(1):e232714.
- Ahyad RA, et al. J Imaging Inform Med. 2025;38(1):346–356.