



A Comparative Evaluation of a Domain-Specific Large Language Model and GPT-4o for Radiology Report Summarization



S. Zheng¹, S. Liu², J. Wang³, X. Cui¹, Z. Ye¹

¹Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin's Clinical Research Center for Cancer, State Key Laboratory of Druggability Evaluation and Systematic Translational Medicine, Tianjin Key Laboratory of Digestive Cancer, Key Laboratory of Cancer Prevention and Therapy, Department of Radiology, Tianjin, China; ²Med Frontier Suppliers & Consultancy UK Limited; ³Tianjin University of Traditional Chinese Medicine, School of Public Health, Tianjin, China; **Correspondence:** zye@tmu.edu.cn

INTRODUCTION

Although the general-purpose large language model (LLM) GPT-4o has shown potential in radiology-related language tasks, it is not yet clear whether its performance in report summarization exceeds that of a domain-specific model developed for this task. This study aimed to compare a domain-specific large language model with GPT-4o in the summarization of radiology reports.

METHODS

A domain-specific large language model was trained using 956,219 CT and MRI reports from our institution. Its performance was first assessed using F1 scores on internal and external validation sets containing 106,247 and 17,091 reports, respectively, from five hospitals. For comparison with GPT-4o, we further selected a balanced subset of 1,800 reports across imaging modalities and anatomical regions. Human reviewers evaluated the outputs of the two models in terms of factual accuracy, coherence of impressions, and considerations related to medical safety. Differences in F1 scores between the models were examined using a t-test.

RESULTS

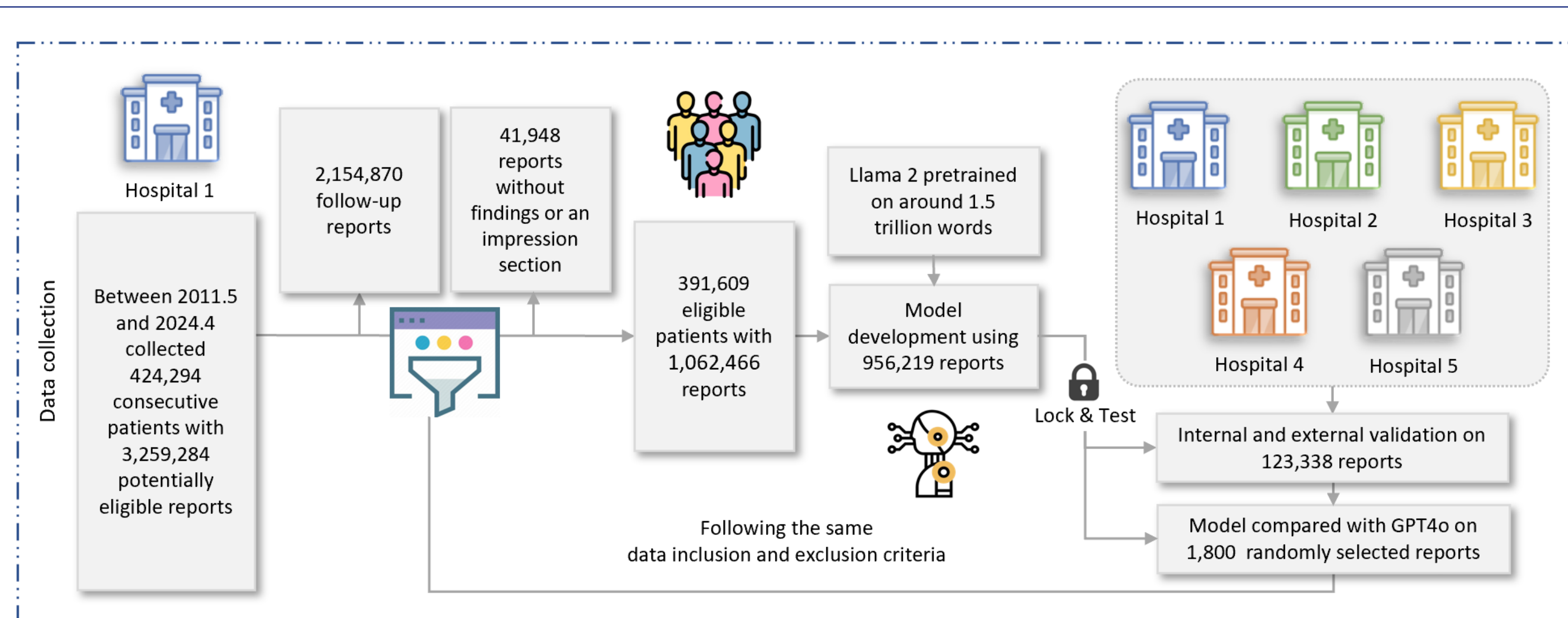


Fig.1 Study flow chart

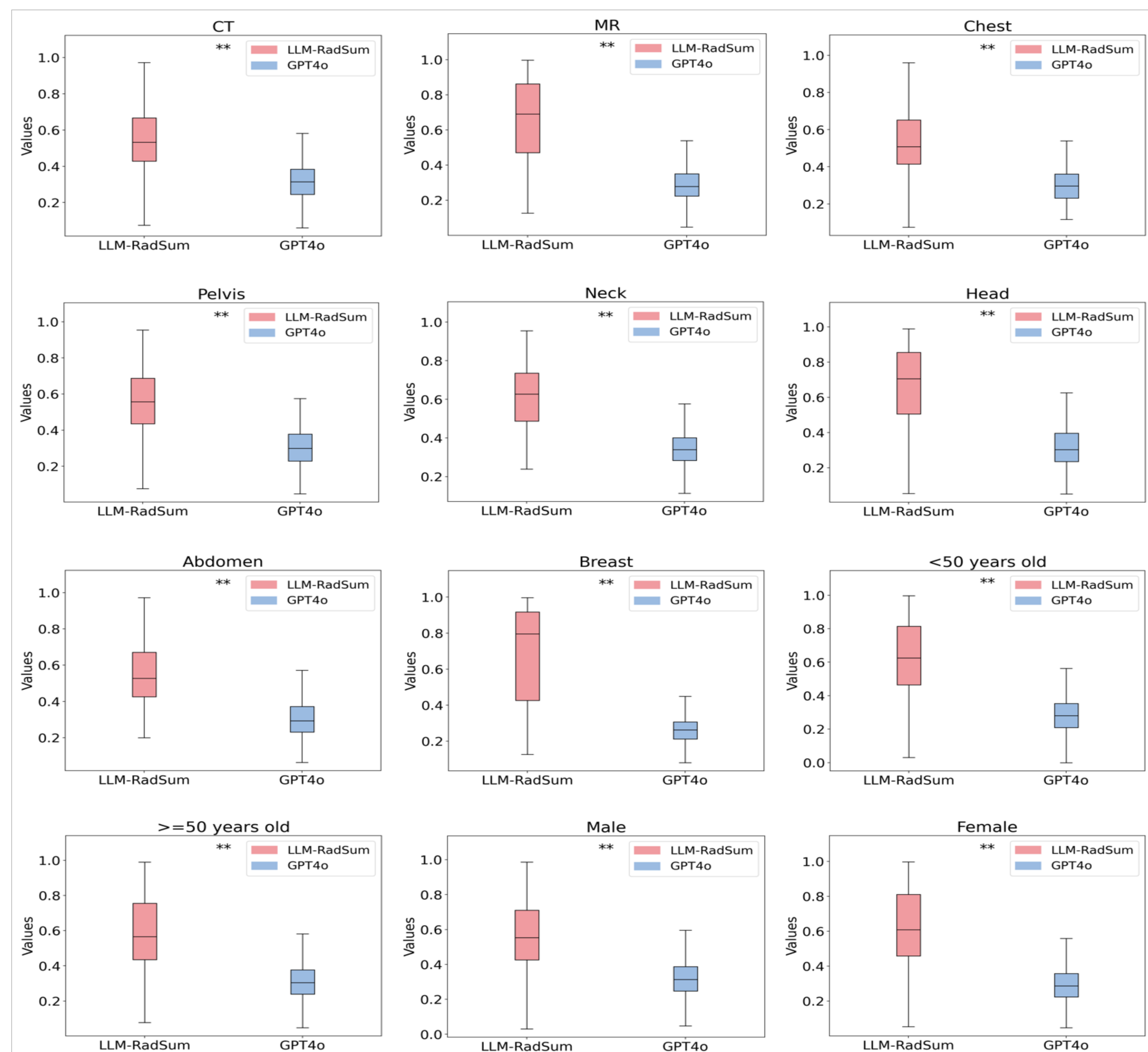


Fig.2 F1 scores of models across different subgroups

RESULTS

The domain-specific model achieved median F1 scores of 0.75 and 0.44 on the internal and external validation sets. According to human assessment, more than 81.5% of its summaries were rated as satisfactory regarding factual correctness, impression clarity, and safety. In contrast, over 40.0% of GPT-4o outputs required minor revisions in these areas. Across anatomical sites, imaging modalities, gender, and age groups, the domain-specific model showed higher F1 scores than GPT-4o (all $p < 0.001$).

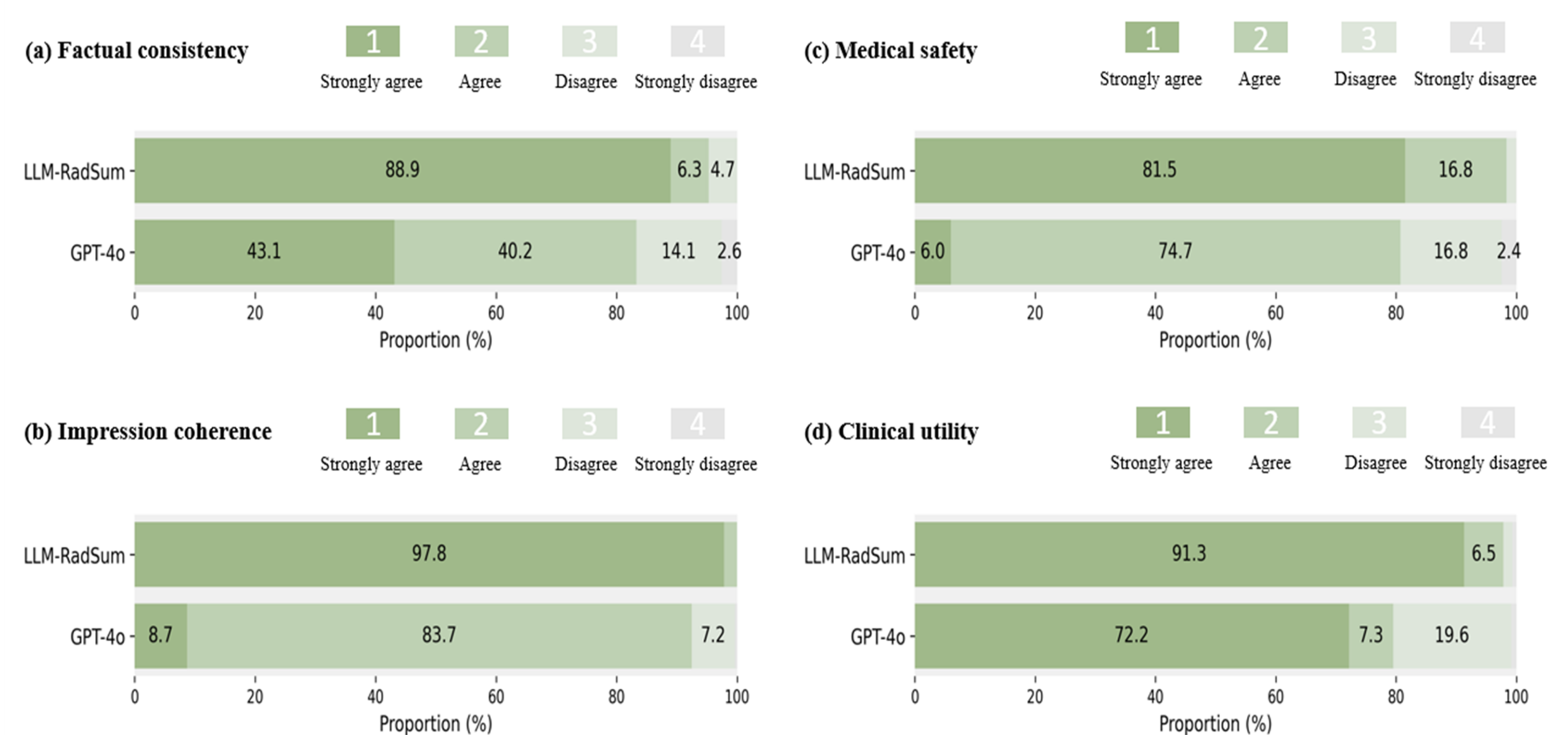


Fig.3 Human evaluation of model performance

DISCUSSION AND CONCLUSION

Actionable Take Away Message

- Domain-specific LLMs may outperform general-purpose models (e.g., GPT-4o) in radiology report summarization, particularly in factual accuracy and clinical safety.
- Training on large-scale, task-specific radiology data enables better alignment with radiologists' reporting patterns and improves summary quality.

Future Direction for Research

- Future work should validate model performance across diverse clinical settings and populations to improve generalizability.
- Integrating domain-specific fine-tuning with real-world clinical deployment and human-in-the-loop evaluation will be critical for safe adoption.

REFERENCES

- Radiology 2023;307(5):e231259.
- Radiology 2024;311(1):e232714.
- Radiology 2023;307(4):e230725.
- Radiology 2025;314(1):e240895.
- Nat Med 2024;30(4):1134–1142.
- Inform Med Unlocked 2024;46:101465.
- IEEE Trans Multimedia 2023;25:845–855.
- Proc AAAI Conf Artif Intell 2024;5034–5042.