

Augmenting Emergency Imaging Decisions: Performance of a Custom Large-Language-Model Decision Support System for SHOs in Time-critical Medical Emergencies

Ebinesh Arulnathan, Praislin Gideon, Shirou Masoodi, Swetha Tula, Matthew Kueh, Rayya Naffa, Sivakumar Manickam, Mohammed Shoaib

Background

- Emergency imaging decisions are time-critical and vulnerable to variability at SHO level.
- A custom ChatGPT model was built using iRefer, clinical guidelines, meta-analyses, and evidence-based recommendations.
- Aim was to compare the performance of SHO alone, custom ChatGPT alone, and SHO + custom ChatGPT for emergency imaging decision support.

Methods

- 50 high-acuity scenarios
- 3 experimental conditions:
 - R1: SHO without AI
 - R2: custom ChatGPT alone
 - R3: SHO + custom ChatGPT
- Questions asked:
 - Q1: need for imaging
 - Q2a: imaging modality
 - Q2b: protocol / contrast / phase
- Single FRCR-certified radiologist reviewer.

Scoring

Q1: 0 or 1
Q2a: 0 / 0.5 / 1
Q2b: 0 / 0.5 / 1
Scenario score = mean of Q1, Q2a, Q2b
Total score summed to /50

Image Descriptions

- Radar chart summarises multidimensional responder performance at a glance.
- Dumbbell plot emphasises progression from baseline (R1) to AI-assisted performance (R3).
- Heatmap highlights domain heterogeneity and visually identifies weakest/strongest areas.
- Forest plot shows effect size magnitude and precision with confidence intervals.

Key Findings

R1 total score
32.5 / 50
Accuracy 65.0%

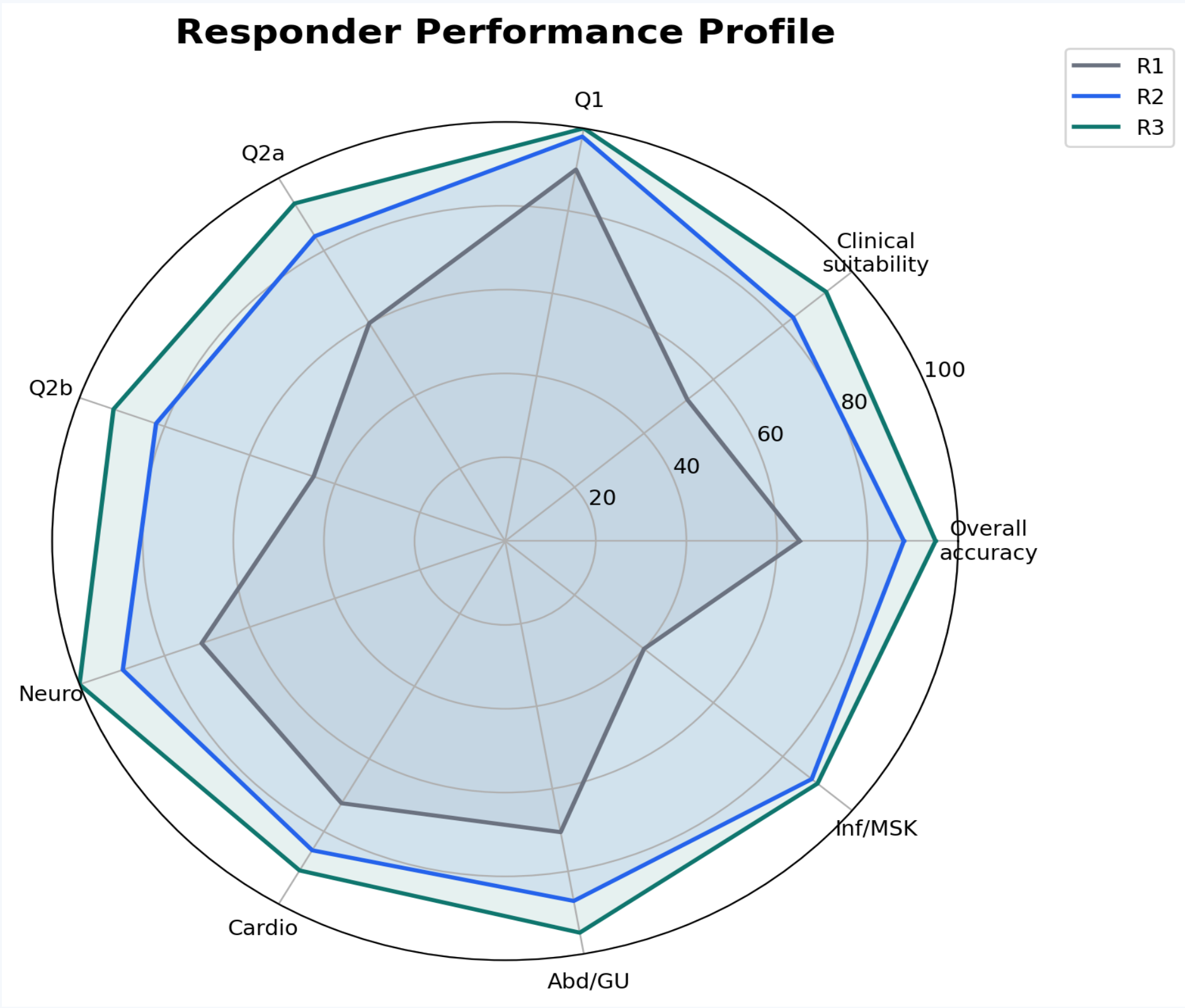
R2 total score
44.0 / 50
Accuracy 88.0%

R3 total score
47.5 / 50
Accuracy 95.0%

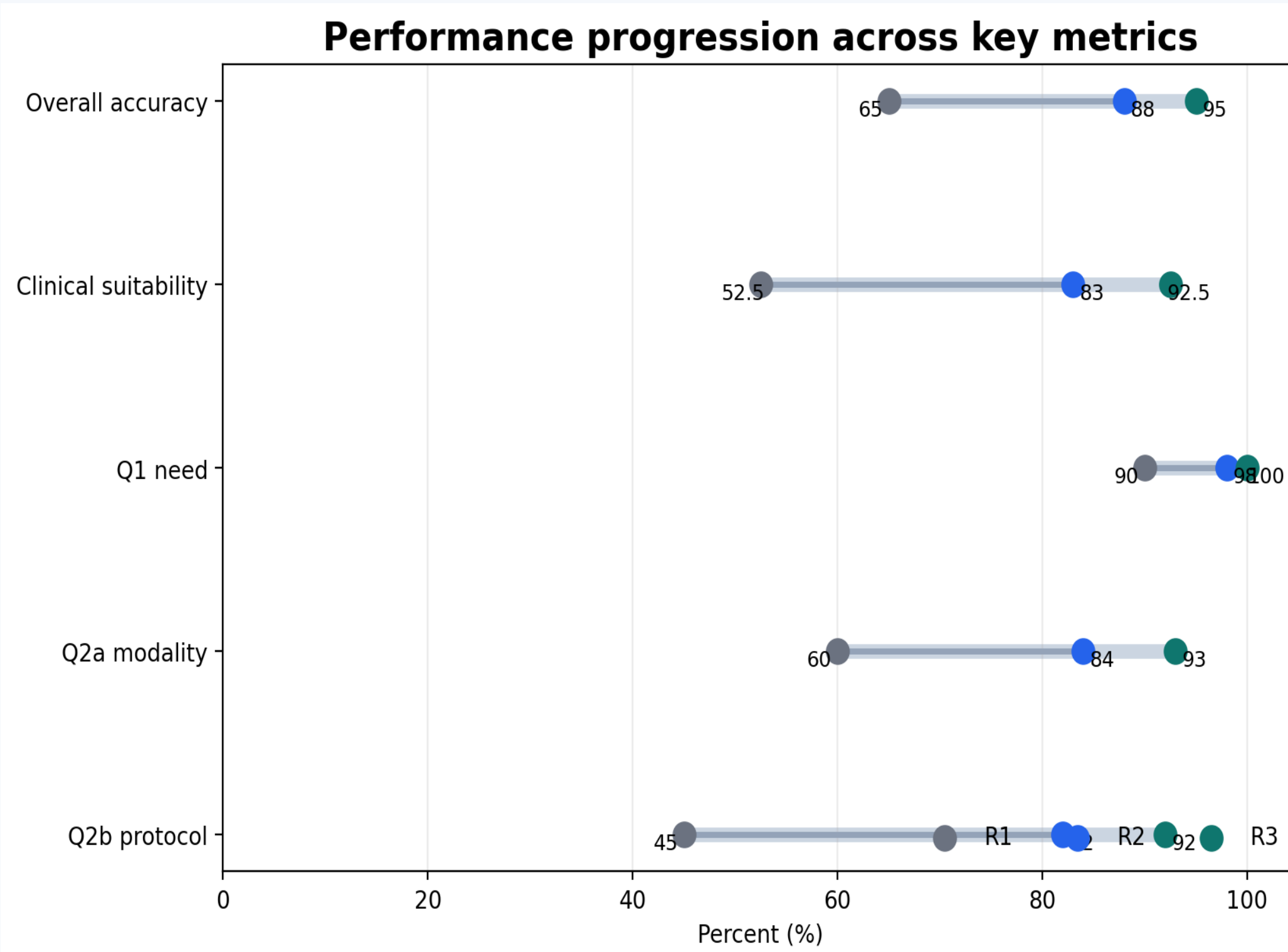
Clinical suitability
52.5% / 83.0% / 92.5%
R1 / R2 / R3

Overall Accuracy
65% / 88% / 95%
R1 / R2 / R3

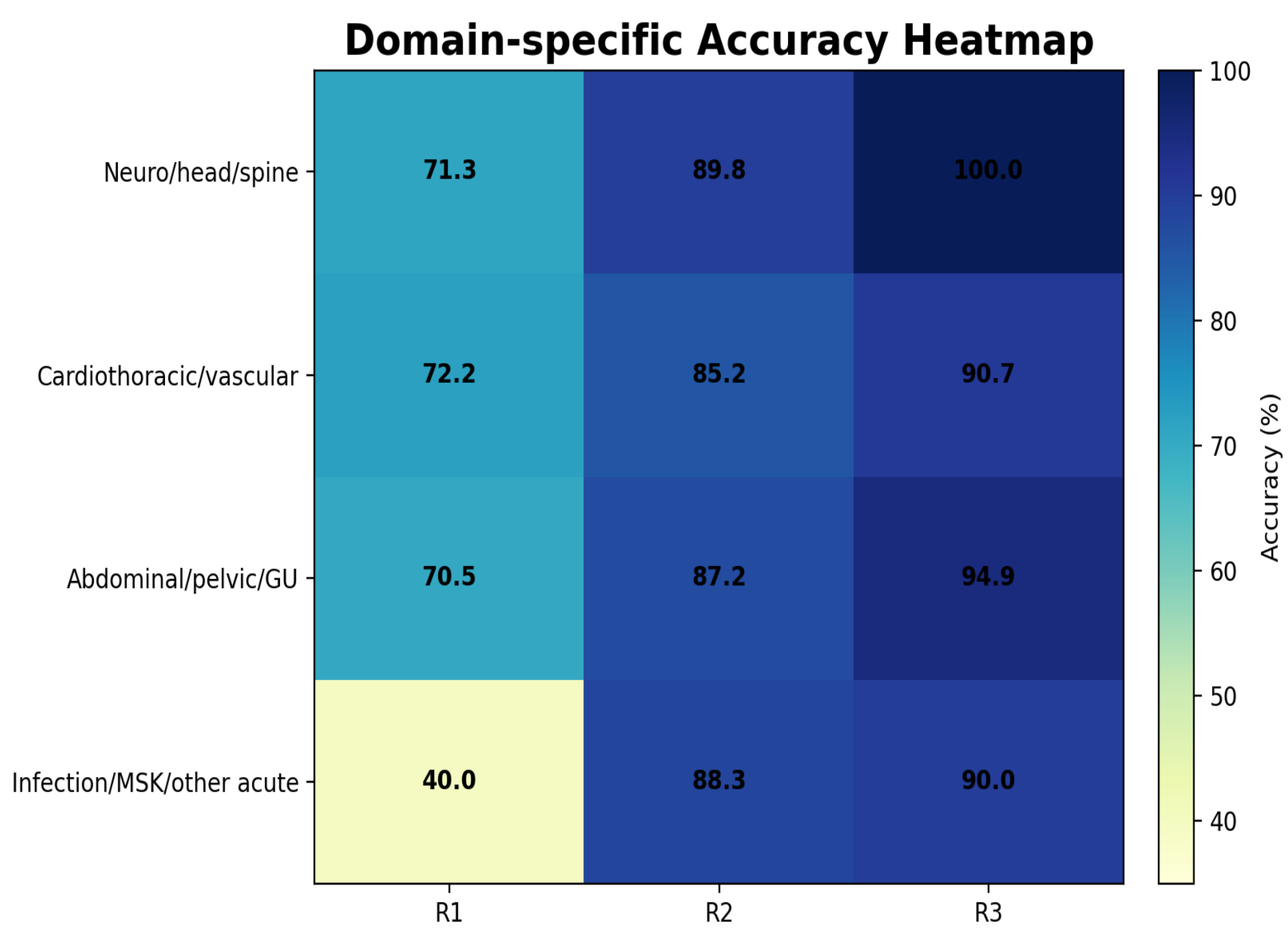
Multidimensional responder profile



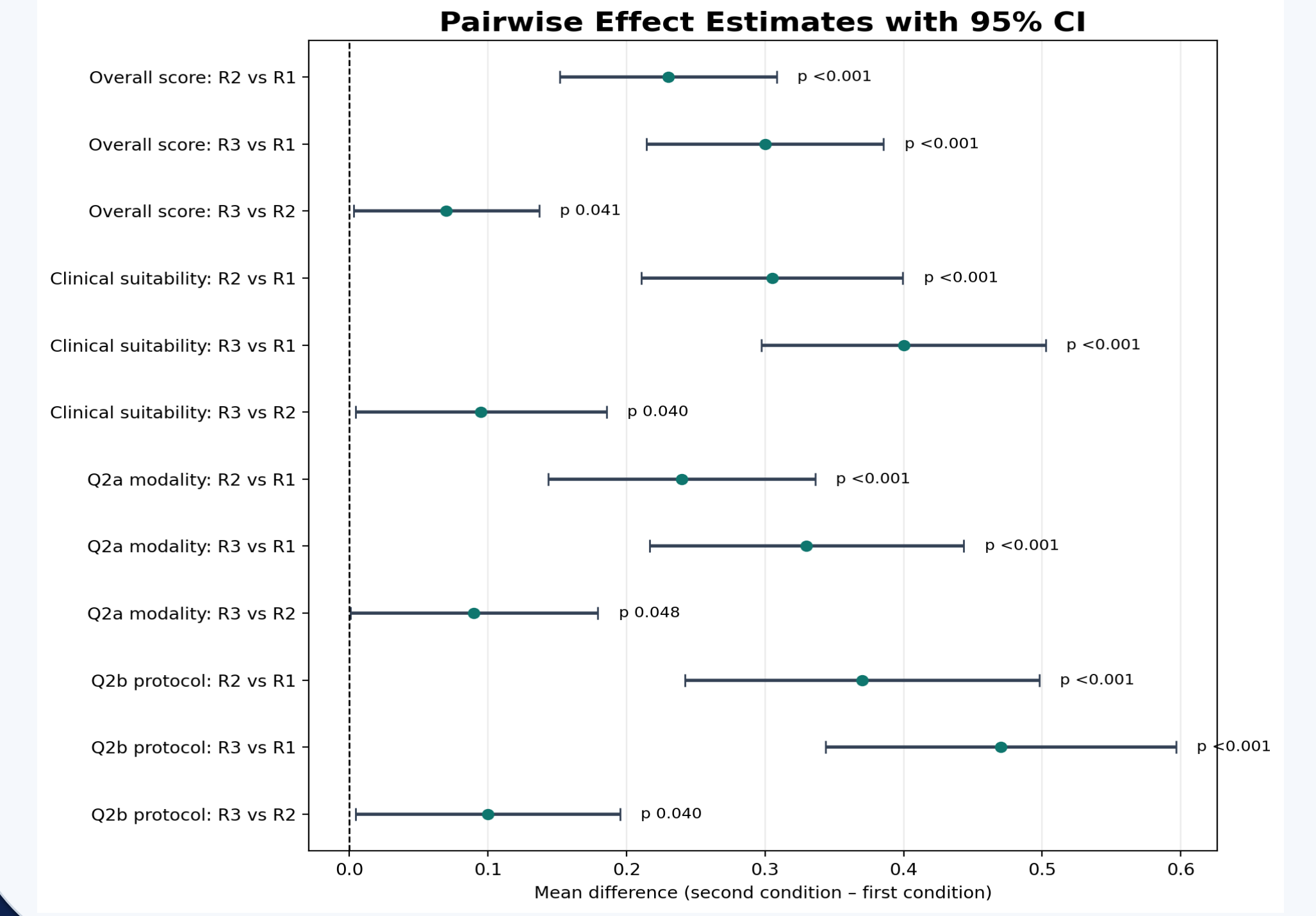
Progression across key metrics



Domain-specific heatmap



Pairwise effect estimates with 95% CI



Statistical highlights

Key Statistical Tests

Overall score: RM ANOVA $F(2,98)=33.26$, $p<0.001$; Friedman $\chi^2(2)=50.72$, $p<0.001$.
Clinical suitability: RM ANOVA $F(2,98)=38.28$, $p<0.001$.
Q1 imaging decision: small but significant omnibus effect (Cochran Q $p=0.015$).
Q2a and Q2b both highly significant across conditions.

Pairwise comparisons

R2 vs R1: +11.5/50 overall score; $p<0.001$.
R3 vs R1: +15.0/50 overall score; $p<0.001$.
R3 vs R2: +3.5/50 overall score; $p=0.041$.
Largest gain: Q2b protocol details 45% $\rightarrow$ 92%.

Interpretation

- R3 was best across all metrics and domains.
- R2 substantially outperformed the unaided SHO.
- AI benefit was greatest for modality and protocol precision rather than simply deciding to image.
- No interobserver agreement analysis was applicable because there was one reviewer.

Discussion

- Intermodel performance:** the custom ChatGPT model alone clearly outperformed the unaided SHO, indicating that a guideline-grounded LLM can deliver more appropriate emergency imaging recommendations.
- Human-AI augmentation:** the best-performing condition was SHO + ChatGPT, supporting a collaborative model in which AI augments, rather than replaces, junior clinical decision making.
- Domain-specific performance:** the largest baseline weakness was in infection/MSK/other acute emergencies; this was also the domain with the greatest improvement.
- Criteria-specific performance:** The most meaningful performance improvement was not just in deciding whether imaging was needed, because Q1 had a ceiling effect. The larger and clinically more relevant improvement was in choosing the correct imaging modality and specifying protocol/contrast/phase details.

Strengths

- Repeated-measures design using the same 50 scenarios.
- Multi-criteria scoring framework.
- Use of both parametric and nonparametric tests.
- Clinically interpretable outputs.

Limitations

- Single-reviewer scoring only.
- Scenario-based design, not real-world deployment.
- Limited sample of 50 scenarios.
- Performance depends on prompt and source curation.

Conclusions

1 Custom ChatGPT significantly improved emergency imaging decision quality versus unaided SHO performance.

2 SHO + custom ChatGPT achieved the highest overall accuracy and clinical suitability.

3 The most meaningful gains were in selecting the correct modality and protocol / contrast details.

4 These findings support further prospective evaluation of LLM-assisted radiology decision support.