

AI-assisted question generation for FRCR Part 1 Physics: Evaluating the role of multi-model LLMs in high-stakes radiology assessment design

Ebinesh Arulnathan, Shirou Masoodi, Zobia Hussain, Rayya Naffa, Srikrishna Harave, Sivakumar Manickam

BACKGROUND

- High-quality MCQs are essential to FRCR Part 1 Physics and high-stakes radiology assessment.
- Manual item writing is time-intensive, costly, and difficult to scale.
- Large language models may accelerate item-bank development, but their output requires careful benchmarking.
- Prior literature suggests frontier models can approach board-style item quality under expert review.

AIMS

- To compare the quality and suitability of FRCR Part 1 Physics questions generated by LLM models.
- To assess overall, domain-specific, criterion-specific, and reviewer-related performance using combined scores from 5 independent reviewers.

METHODS

1
Source

105 total questions;
35 per model; ChatGPT, Gemini
and Perplexity

2
Sampling
5 per source
× 7 domains

Matter & Radiation;
Radiography & Fluoroscopy;
Radionuclide Imaging; Radiation
Safety; CT; MRI; Ultrasound

3
Anonymisation

Source-anonymised with
random IDs
5 independent qualified
reviewers

4
Scoring
7 criteria

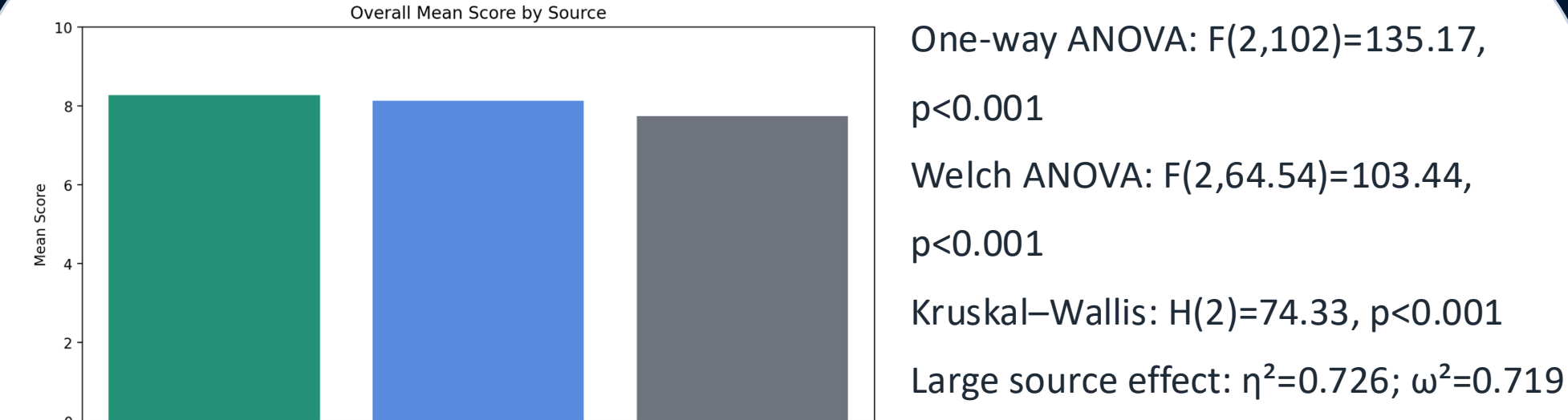
Clarity; Relevance; Suitability;
Technical Accuracy; Difficulty;
Adequacy of Explanation;
Reliability of Rationale

5
Analysis

ANOVA, Welch ANOVA,
Kruskal–Wallis, Friedman,
pairwise tests, FDR
correction

RESULTS

OVERALL PERFORMANCE

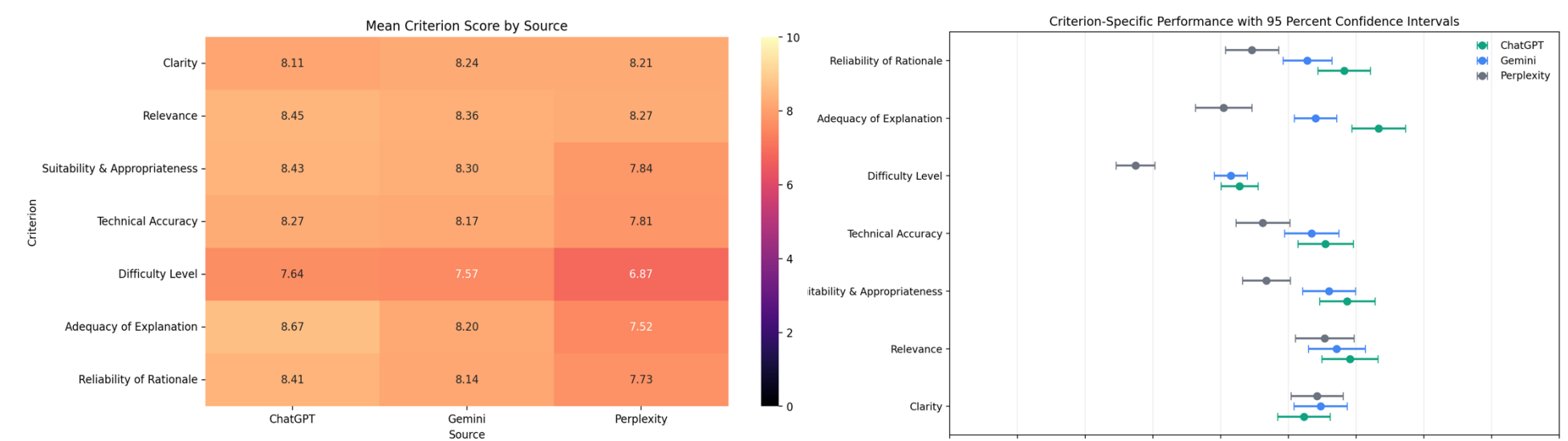


Pairwise ranking: ChatGPT > Gemini > Perplexity

Top-line finding

ChatGPT achieved the highest overall score and led all 7 domains; Gemini was close, while Perplexity underperformed mainly in depth and rationale quality.

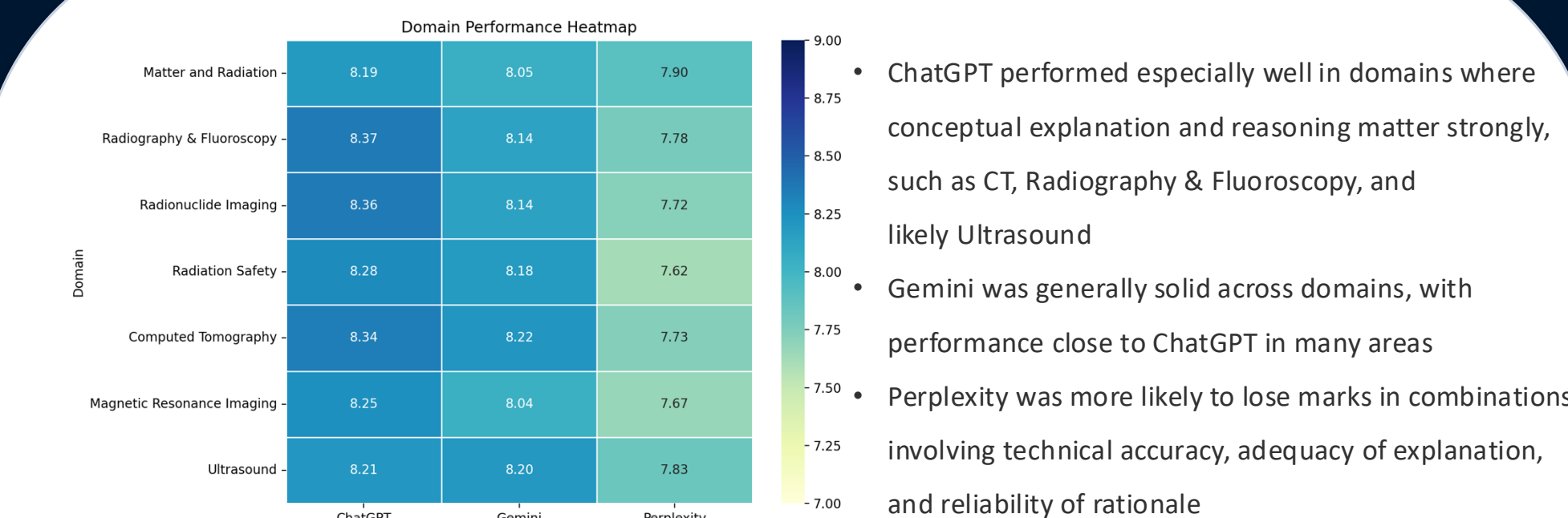
CRITERION-SPECIFIC PERFORMANCE



Criterion insights

- Gemini was highest for Clarity.
- ChatGPT led 6/7 criteria.
- Difficulty Level was the weakest global criterion.
- Friedman $\chi^2(6)=285.47$, $p<0.001$; Kendall's $W=0.453$.

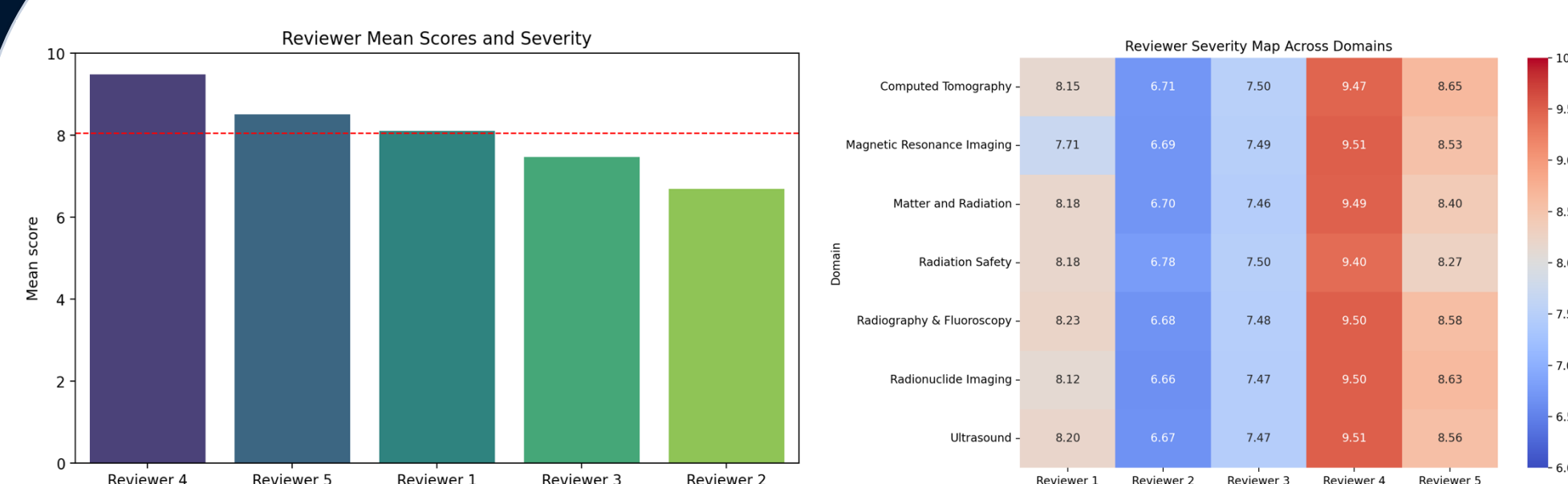
DOMAIN-SPECIFIC PERFORMANCE



Domain interpretation

- Source main effect was strong.
- Domain main effect was small and non-significant.
- ChatGPT ranked first in all 7 domains.
- Gemini was strongest in CT (8.22) and Ultrasound (8.20).

REVIEWER VARIABILITY



Reviewer variability

- Reviewer stringency varied substantially: ANOVA $F(4,520)=568.30$, $p<0.001$; Kruskal–Wallis $H(4)=434.38$, $p<0.001$.
- Combined reviewer means were therefore used for primary analysis, supporting multi-reviewer validation.

DISCUSSION

Inter-model

ChatGPT was highest overall, Gemini a strong second, and Perplexity lower mainly in explanation depth, rationale reliability, and difficulty calibration.

Intra-model

ChatGPT was the most stable across domains (8.19–8.36); Gemini was consistent (8.05–8.22); Perplexity varied more (7.62–7.90).

Criterion lens

Models were closest in clarity and relevance; differences were largest in adequacy of explanation and reliability of rationale.

Reviewer lens

Marked reviewer stringency heterogeneity reinforces the need for multiple reviewers rather than single-expert approval.

COMPARISON WITH LITERATURE

- Our findings align with Mistry et al. (Academic Radiology, 2024), who found that a frontier LLM (GPT-4) approached board-style MCQ quality under expert review, whereas a weaker model underperformed.
- GPT-4 scored ~9.8–9.9 across quality criteria; Llama 2 was weaker, especially for adequacy of rationale (7.2), with 69% answer accuracy versus 100% for GPT-4.
- The current study extends this literature by directly comparing three contemporary models, focusing specifically on FRCR Part 1 Physics, using 7 domains, 5 independent reviewers, and robust parametric plus nonparametric testing.
- The shared message is that powerful LLMs can assist question drafting, but human review remains essential.

EDUCATIONAL IMPLICATIONS

AI draft

Expert review

Pilot test

Deploy

AI can support scalable question-bank drafting but does not replace human item writing and governance.

The key bottleneck is difficulty calibration and verification of explanation/rationale fidelity.

CONCLUSIONS

- ChatGPT demonstrated the strongest overall performance and led all domains.
- Gemini was a credible second, with the clearest wording but lower explanation depth.
- Perplexity underperformed relative to both competitors.
- AI-assisted item generation is promising, but expert validation remains essential before high-stakes use.

KEY REFERENCE

Mistry NP, Saeed H, Rafique S, et al. Large Language Models as Tools to Generate Radiology Board-Style Multiple-Choice Questions. Academic Radiology. 2024;31:3872–3878.