Artificial Intelligence in Periprosthetic Joint Infection: Are All Large Language Models Equally Reliable?

R. Rojas-Sayol[1], O. Pujol[2], R. Ferrer[3], M. Vicente[2], M. Monfort-Mira[4], M. Sabater-Martos[5], D. Pérez-Prieto[6], A. Coelho[1]

[1] Hospital Universitari Parc Taulí; [2] Hospital Universitari Vall d'Hebron; [3] Hospital Sant Rafael; [4] Clínic Barcelona Hospital; [5] Hospital Universitari Joan XXIII; [6] Hospital del Mar. Orthopaedic Surgery Departments, Spain.

Introduction & Aim

- Periprosthetic joint infection (PJI) remains a complex, high-impact complication requiring accurate and evidence-based clinical decisions.
- Large language models (LLMs) are increasingly used in medicine, but their reliability for expert, consensus-based PJI questions remains uncertain.
- Aim: to evaluate and compare five LLMs using clinical questions derived from the 2025 International Consensus Meeting (ICM).

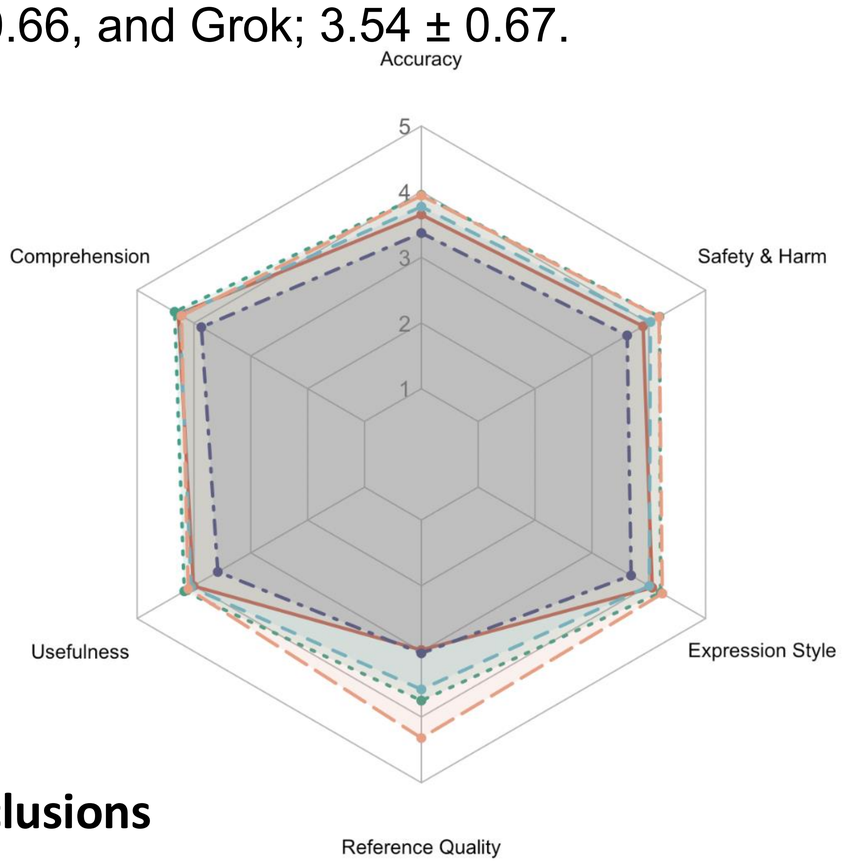
Methods

- Cross-sectional comparative study using 20 clinically relevant ICM 2025 questions covering PJI prevention, diagnosis and management.
- Five LLMs were assessed: ChatGPT, Claude, Grok, Gemini and OpenEvidence.
- Five orthopedic infection specialists independently performed blinded evaluations. Responses were scored from 1 to 5 across six adapted QUEST domains: accuracy, comprehensiveness, clinical utility, reference quality, writing style and safety.
- Statistical analysis: Friedman tests with post-hoc paired comparisons and mixed-effects models

20 questions • 5 LLMs • 5 reviewers

Results

- Significant differences in overall performance were observed ($p < 0.001$). Mean ± SD QUEST scores: OpenEvidence; 4.17 ± 0.56, ChatGPT; 4.10 ± 0.58, Claude; 3.95 ± 0.57, Gemini; 3.87 ± 0.66, and Grok; 3.54 ± 0.67.
- ChatGPT achieved the highest Accuracy (3.95) and Comprehensiveness (4.34).
- OpenEvidence achieved the highest Reference Quality (4.32) and received the most best-answer nominations (31%).
- Overall inter-rater reliability was moderate (ICC = 0.635). Differences were especially relevant in Accuracy and Reference Quality



Conclusions

- LLMs demonstrated strong performance in answering expert-level PJI questions, although meaningful differences were observed between models, particularly in accuracy and reference quality. This variability highlights limitations in high-stakes clinical settings.
- LLMs may serve as useful adjuncts in orthopaedic infection practice, but their outputs require careful interpretation and should not replace expert clinical judgement.