Supplementary Table S2.

Holm-adjusted pairwise comparisons (p-values) between AI models across evaluation domains

Comparison	Clarity	Clinical Utility	Factual Accuracy	Relevance	Structure
EAU Guidelines Bot vs ChatGPT-5	0.236	0.034	0.267	0.930	0.020
EAU Guidelines Bot vs Gemini 2.5 Pro	0.016	0.041	0.348	0.915	1.000
EAU Guidelines Bot vs Copilot – Smart GPT-5	0.062	0.016	0.013	0.016	0.017
EAU Guidelines Bot vs Perplexity Pro	0.236	0.016	0.012	0.016	0.016
ChatGPT-5 vs Gemini 2.5 Pro	0.056	0.285	1.000	0.915	0.032
ChatGPT-5 vs Copilot – Smart GPT-5	0.016	0.063	0.014	0.016	1.000
ChatGPT-5 vs Perplexity Pro	0.022	0.034	0.013	0.016	1.000
Gemini 2.5 Pro vs Copilot – Smart GPT-5	0.016	0.034	0.014	0.016	0.016
Gemini 2.5 Pro vs Perplexity Pro	0.016	0.016	0.013	0.019	0.016
Copilot – Smart GPT-5 vs Perplexity Pro	0.062	0.034	0.229	0.296	0.341