Supplementary Table S4. Test–retest consistency of AI responses across specialty-role sessions

A. Single-response clinical decision items

AI role	Item	Model	Exact agreement (n/N)	Exact agreement (%)	Cohen's κ
Orthodontist	S1	ChatGPT-5.6	10/12	83.3	0.760
		Gemini Pro	8/12	66.7	0.600
		Claude Opus 4.8 High	6/12	50.0	0.400
Orthodontist	S2	ChatGPT-5.6	9/12	75.0	0.679
		Gemini Pro	10/12	83.3	0.848
		Claude Opus 4.8 High	8/12	66.7	0.684
Orthodontist	S3	ChatGPT-5.6	10/12	83.3	0.273
		Gemini Pro	8/12	66.7	0.351
		Claude Opus 4.8 High	11/12	91.7	0.838
Orthodontist	S4	ChatGPT-5.6	6/12	50.0	0.000
		Gemini Pro	7/12	58.3	0.118
		Claude Opus 4.8 High	7/12	58.3	0.000
Orthodontist	O1a	ChatGPT-5.6	10/12	83.3	0.625
		Gemini Pro	12/12	100.0	1.000
		Claude Opus 4.8 High	11/12	91.7	0.824
Orthodontist	O2	ChatGPT-5.6	9/12	75.0	0.769
		Gemini Pro	6/12	50.0	0.548
		Claude Opus 4.8 High	7/12	58.3	0.523
Orthodontist	Pooled single-response items	ChatGPT-5.6	54/72	75.0	—
		Gemini Pro	51/72	70.8	—
		Claude Opus 4.8 High	50/72	69.4	—
Periodontist	S1	ChatGPT-5.6	11/12	91.7	0.854
		Gemini Pro	10/12	83.3	0.778
		Claude Opus 4.8 High	9/12	75.0	0.647
Periodontist	S2	ChatGPT-5.6	7/12	58.3	0.434

		Gemini Pro	11/12	91.7	0.918
		Claude Opus 4.8 High	8/12	66.7	0.662
		ChatGPT-5.6	9/12	75.0	0.000
Periodontist	S3	Gemini Pro	9/12	75.0	0.561
		Claude Opus 4.8 High	10/12	83.3	0.707
		ChatGPT-5.6	5/12	41.7	0.000
Periodontist	S4	Gemini Pro	9/12	75.0	0.500
		Claude Opus 4.8 High	8/12	66.7	0.000
		ChatGPT-5.6	10/12	83.3	0.750
Periodontist	P1	Gemini Pro	10/12	83.3	0.850
		Claude Opus 4.8 High	12/12	100.0	1.000
		ChatGPT-5.6	11/12	91.7	0.824
Periodontist	P2a	Gemini Pro	12/12	100.0	1.000
		Claude Opus 4.8 High	12/12	100.0	1.000
		ChatGPT-5.6	3/4	75.0	0.000
Periodontist	P2c	Gemini Pro	4/4	100.0	—
		Claude Opus 4.8 High	3/4	75.0	0.000
		ChatGPT-5.6	56/76	73.7	—
Periodontist	Pooled single-response items	Gemini Pro	65/76	85.5	—
		Claude Opus 4.8 High	62/76	81.6	—

B. Multiple-response clinical decision items

AI role	Item	Model	Eligible paired cases (N)	Exact response-set agreement (n/N)	Exact response-set agreement (%)	Mean Jaccard similarity
		ChatGPT-5.6	7	2/7	28.6	0.74
Orthodontist	O1b	Gemini Pro	7	1/7	14.3	0.70
		Claude Opus 4.8 High	7	1/7	14.3	0.44
		ChatGPT-5.6	3	1/3	33.3	0.50
Periodontist	P2b	Gemini Pro	4	4/4	100.0	1.00
		Claude Opus 4.8 High	4	2/4	50.0	0.75

*The 12-case subset comprised cases 3, 8, 11, 17, 18, 23, 25, 31, 35, 39, 44, and 48. Each original AI response was compared with one repeat response. Exact agreement denotes identical responses across assessment occasions. Linear-weighted Cohen's κ was used for ordinal items (S1, S2, S4, O2, P1); unweighted κ was used for nominal/binary items (S3, O1a, P2a, P2c). For multiple-response items (O1b, P2b), exact response-set agreement requires identical selected-option sets; mean Jaccard similarity measures overlap between sets. Conditionally applicable items were analyzed only for eligible paired responses. A dash indicates that κ was not estimable or was not calculated (pooled rows). Pooled exact agreement is reported descriptively as n/N and percentages; no confidence intervals are presented because multiple items from the same case are not independent.