Supplementary Table S1. Agreement of AI models with specialty-specific expert-based references for shared items (S1–S4) in the sensitivity analysis.

Item	Model	Orthodontist session			Periodontist session		
		n	Exact agreement, n/N; % (95% CI)	Cohen's κ (95% CI)	n	Exact agreement, n/N; % (95% CI)	Cohen's κ (95% CI)
S1	ChatGPT-5.6	48	41/48; 85.4 (72.8–92.8)	0.838 (0.728–0.933)	48	27/48; 56.3 (42.3–69.3)	0.500 (0.303–0.664)
	Gemini Pro	48	39/48; 81.3 (68.1–89.8)	0.797 (0.668–0.907)	48	30/48; 62.5 (48.4–74.8)	0.547 (0.343–0.712)
	Claude Opus 4.8 High	48	37/48; 77.1 (63.5–86.7)	0.754 (0.613–0.873)	48	31/48; 64.6 (50.4–76.6)	0.584 (0.377–0.745)
S2	ChatGPT-5.6	48	31/48; 64.6 (50.4–76.6)	0.611 (0.444–0.745)	48	32/48; 66.7 (52.5–78.3)	0.612 (0.469–0.736)
	Gemini Pro	48	36/48; 75.0 (61.2–85.1)	0.759 (0.618–0.871)	48	35/48; 72.9 (59.0–83.4)	0.725 (0.562–0.853)
	Claude Opus 4.8 High	48	29/48; 60.4 (46.3–73.0)	0.591 (0.452–0.716)	48	33/48; 68.8 (54.7–80.1)	0.676 (0.556–0.788)
S3	ChatGPT-5.6	47	27/47; 57.4 (43.3–70.5)	0.311 (0.155–0.479)	46	15/46; 32.6 (20.9–47.0)	0.099 (0.024–0.187)
	Gemini Pro	47	32/47; 68.1 (53.8–79.6)	0.500 (0.313–0.680)	46	20/46; 43.5 (30.2–57.8)	0.259 (0.109–0.418)
	Claude Opus 4.8 High	47	30/47; 63.8 (49.5–76.0)	0.434 (0.256–0.614)	46	24/46; 52.2 (38.1–65.9)	0.348 (0.204–0.507)
S4	ChatGPT-5.6	48	26/48; 54.2 (40.3–67.4)	—	48	10/48; 20.8 (11.7–34.3)	—
	Gemini Pro	48	31/48; 64.6 (50.4–76.6)	0.299 (0.032–0.548)	48	27/48; 56.3 (42.3–69.3)	0.058 (−0.198–0.312)
	Claude Opus 4.8 High	48	17/48; 35.4 (23.4–49.6)	0.030 (0.000–0.101)	48	11/48; 22.9 (13.3–36.5)	0.037 (0.000–0.105)

Note: This sensitivity analysis used specialty-specific expert references derived separately from the independent assessments of the three orthodontists and the three periodontists. For ordinal items (S1, S2, and S4), the reference was defined as the median rating of the three experts; for the nominal item (S3), it was defined as the modal category. The resulting denominators were 48 for S1, S2, and S4 in both specialty perspectives. For S3, the denominators were 47 for the orthodontist panel and 46 for the periodontist panel because no unique mode could be established when all three experts selected different categories (one orthodontist-panel case and two periodontist-panel cases were excluded). Each AI output was compared with the reference derived from the corresponding specialty. Linear-weighted Cohen's κ was used for ordinal items, and unweighted Cohen's κ was used for S3. Exact-agreement 95% confidence intervals were calculated using the Wilson method; 95% confidence intervals for κ were estimated using case-level bootstrap resampling with 5,000 replicates. An em dash indicates that κ could not be estimated because ChatGPT-5.6 gave the same response in all cases. CI, confidence interval.