

Supplementary

Appendix 1

Available at <https://cdn.amegroups.cn/static/public/JMAI-2026-0104-Appendix1.pdf>

Table S1 Inter-specialist variability and reliability of the three-specialist reference standard by landmark

Landmark	Mean deviation (SD), mm	Maximum deviation, mm	ICC(2,3), X	ICC(2,3), Y
S	1.212 (0.857)	2.976	0.983	0.974
N	1.192 (0.781)	2.571	0.992	0.992
A	1.226 (0.797)	2.534	0.992	0.993
B	1.210 (0.906)	2.983	0.991	0.991
UIT	1.128 (0.853)	2.850	0.987	0.995
UIA	1.334 (0.842)	3.326	0.989	0.990
LIT	1.403 (0.869)	2.747	0.992	0.987
LIA	1.261 (0.819)	2.610	0.991	0.991

Deviation was calculated as the Euclidean distance between each specialist coordinate and the corresponding three-specialist mean coordinate and was converted to millimetres using image-specific calibration. Values are based on 33 specialist observations per landmark (3 specialists × 11 images). ICC(2,3) denotes a two-way random-effects, absolute-agreement, average-measures intraclass correlation coefficient calculated separately for the calibrated x and y coordinates. S, Sella; N, Nasion; A, A point; B, B point; UIT, upper incisor tip; UIA, upper incisor apex; LIT, lower incisor tip; LIA, lower incisor apex; SD, standard deviation.

Table S2 Descriptive comparison of unmodified AI-generated landmark predictions and students' final AI-assisted landmark positions

Landmark	Raw AI mean (SD), mm	Final AI-assisted mean (SD), mm	Raw AI median, mm	Final AI-assisted median, mm	Raw AI SDR ≤ 2 mm	Final AI-assisted SDR ≤ 2 mm	Raw AI SDR ≤ 4 mm	Final AI-assisted SDR ≤ 4 mm
Overall	1.431 (1.157)	1.826 (0.926)	1.072	1.733	79.5%	62.9%	93.2%	97.1%
S	0.967 (0.565)	1.646 (0.733)	0.907	1.618	100.0%	67.4%	100.0%	100.0%
N	1.556 (1.402)	1.530 (0.568)	0.975	1.552	81.8%	76.9%	81.8%	100.0%
A	2.704 (1.349)	2.336 (1.180)	2.781	2.168	36.4%	43.6%	81.8%	90.9%
B	0.910 (0.455)	1.775 (0.742)	0.958	1.780	100.0%	59.8%	100.0%	100.0%
UIT	0.754 (0.324)	1.550 (0.659)	0.666	1.595	100.0%	75.0%	100.0%	100.0%
UIA	2.331 (1.459)	2.139 (1.243)	1.880	1.911	54.5%	54.2%	81.8%	91.3%
LIT	0.853 (0.584)	1.580 (0.655)	0.583	1.604	90.9%	72.7%	100.0%	100.0%
LIA	1.372 (0.817)	2.053 (1.000)	1.538	1.937	72.7%	53.8%	100.0%	94.3%

Raw AI results were calculated from one deterministic prediction for each of eight anatomical landmarks across 11 cephalograms, giving 88 unique image–landmark predictions overall and 11 predictions per landmark. Final AI-assisted results were calculated from the saved positions of 24 students across the same images and landmarks, giving 2,112 observations overall and 264 observations per landmark. Errors represent Euclidean distances from the three-specialist reference standard, converted to millimetres using the same image-specific calibration factors. The comparison is descriptive because the raw AI and final student datasets have different replication structures; no inferential testing was performed. SD, standard deviation; SDR, success detection rate; AI, artificial intelligence; S, sella; N, nasion; A, A Point; B, B Point; UIT, upper incisor tip; UIA, upper incisor apex; LIT, lower incisor tip; LIA, lower incisor apex.