Online Resource 4 — Supplementary statistical results (Tables S1–S8)

Quality, guideline concordance, and readability of large language model responses to patient questions on robotic surgery in gynecology and gynecologic oncology: a bilingual Turkish–English cross-sectional study | Journal of Robotic Surgery | Authors: Zınar Serhanoğlu; Sema Karakaş; Burcu Akay Serhanoğlu; Fulya Gökdağlı Sağır; Özgür Aslan; Basri Çakıroğlu | Corresponding author: dr.zinarserhanoglu@gmail.com

Data sources: two independent clinical rating workbooks matched to the 252-response coordinator key. Scores were retained without inventing or imputing missing clinician evaluations.

Table S1a. Inter-rater reliability

Outcome	ICC(A,1)	95% bootstrap CI	ICC(A,2)	95% bootstrap CI
QAMAI total	0.728	0.639 to 0.795	0.842	0.780 to 0.886
Independent accuracy (1–5)	0.433	0.267 to 0.572	0.604	0.421 to 0.727
JAMA total	0.727	0.647 to 0.790	0.842	0.786 to 0.883

Bootstrap confidence intervals use 5,000 resamples of matched response–rater pairs. The prespecified ICC(A,1) acceptability threshold was 0.75. Averages used in inferential analyses were prespecified in Methods.

Table S1b. Agreement on individual QAMAI items

QAMAI item	Quadratic weighted κ	Exact agreement (%)
Q1 Accuracy	0.481	76.2
Q2 Clarity	0.363	81.7
Q3 Relevance	0.035	69.8
Q4 Completeness	0.769	87.3
Q5 Sources	0.723	75.4
Q6 Usefulness	0.844	90.5

Table S1c. Binary agreement

Binary item	Cohen κ	Agreement (%)	Both yes	Either yes
Harm	0.665	99.6	1	2
B1 Alternative route	0.811	95.6	213	224
B2 Benefit overstatement	1.000	100.0	1	1
B3 Guideline discordance	0.798	99.6	2	3
B4 Surgeon/centre experience	0.857	92.9	124	142

Cronbach α: S.K. = 0.357; B.A.S. = 0.286. κ may be unstable for rare positive outcomes or strongly skewed scores.

Table S2. Secondary language comparisons at question level (42 matched pairs)

Outcome	TR − EN	95% CI	Test	p	Holm p
Q1 Accuracy	-0.083	-0.172 to 0.005	Wilcoxon	0.206	0.617
Q2 Clarity	+0.143	0.044 to 0.241	Wilcoxon	0.009	0.044
Q3 Relevance	-0.004	-0.081 to 0.073	Wilcoxon	0.894	0.894
Q4 Completeness	-0.369	-0.473 to -0.265	Wilcoxon	7e-07	4.2e-06
Q5 Sources	-0.706	-0.904 to -0.508	paired t	8.4e-09	5.9e-08
Q6 Usefulness	+0.075	-0.065 to 0.216	paired t	0.285	0.617
Independent accuracy (1–5)	-0.091	-0.194 to 0.011	Wilcoxon	0.083	0.333
JAMA total	-0.131	-0.277 to 0.015	paired t	0.077	—

Holm family: Q1–Q6 and independent accuracy (seven comparisons). JAMA is presented separately without an adjusted p value. Paired mean CIs are descriptive even when the primary test was Wilcoxon. Language-specific readability scales are intentionally not contrasted across languages.

Table S3. Liberal rule sensitivity (at least one rater yes)

Binary outcome	TR yes	EN yes	TR only	EN only	McNemar p	Holm p
Harm	1/126	1/126	1	1	1.000	1.000
B1 Alternative route	113/126	111/126	10	8	0.815	1.000
B2 Benefit overstatement	1/126	0/126	1	0	1.000	1.000
B3 Guideline discordance	2/126	1/126	2	1	1.000	1.000
B4 Surgeon/centre experience	63/126	79/126	13	29	0.014	—

Table S4. Binary outcomes compared between chatbots within language

Language	Item	ChatGPT / Gemini / Claude positives	Cochran Q	p	Holm p
TR	Harm	0/1/0	2.00	0.368	0.736
TR	B1 Alternative route	35/34/40	5.17	0.076	0.302
TR	B2 Benefit overstatement	0/1/0	2.00	0.368	0.736
TR	B3 Guideline discordance	0/2/0	4.00	0.135	0.406
TR	B4 Surgeon/centre experience	17/13/24	9.79	0.007	—
EN	Harm	0/0/0	NA	NA	—
EN	B1 Alternative route	35/35/34	0.13	0.936	0.936
EN	B2 Benefit overstatement	0/0/0	NA	NA	—
EN	B3 Guideline discordance	0/0/0	NA	NA	—
EN	B4 Surgeon/centre experience	25/18/27	5.36	0.069	—

Holm correction is shown across Harm and B1–B3 within each language; B4 is exploratory and shown with its unadjusted omnibus p. NA indicates degenerate comparisons with no positive responses, for which Cochran Q is not identifiable.

Exploratory Turkish B4 post hoc McNemar comparisons (Holm corrected among three pairs): ChatGPT vs Gemini adjusted p = 0.344; ChatGPT vs Claude adjusted p = 0.185; Gemini vs Claude adjusted p = 0.022.

Table S5. Exploratory language differences by clinical domain

Domain	Pairs	TR − EN	95% CI	p	Holm p
A1	7	-0.833	-1.894 to 0.227	0.103	0.206
A2	7	-1.167	-1.783 to -0.550	0.004	0.021
A3	7	-1.048	-1.842 to -0.253	0.018	0.072
A4	7	-0.929	-1.911 to 0.054	0.060	0.180
A5	7	-0.310	-1.781 to 1.162	0.625	0.625
A6	7	-1.381	-2.278 to -0.483	0.009	0.047

Domains: A1 benign hysterectomy; A2 myomectomy/endometriosis; A3 sacrocolpopexy; A4 endometrial cancer; A5 cervical cancer; A6 ovarian cancer and risk-reducing surgery. Seven questions per domain, with each question score averaged across three chatbots.

Table S6. Within-language readability tests

Index	Language	Friedman χ^2	p	Holm p
atesman	TR	53.29	2.7e-12	2.7e-12
bezirci_yilmaz	TR	76.19	2.9e-17	5.7e-17
flesch_reading_ease	EN	65.76	5.2e-15	1e-14
flesch_kincaid_grade	EN	57.57	3.2e-13	3.2e-13

Readability values for 252 responses were obtained from the coordinator workbook. The language-specific formulas were not independently recalculated for this supplementary analysis. No English response had a Flesch–Kincaid grade level of 6 or lower.

atesman post hoc Wilcoxon (Holm within three model pairs): ChatGPT vs Gemini adjusted p = 9.5e-12; ChatGPT vs Claude adjusted p = 0.007; Gemini vs Claude adjusted p = 8e-11.

bezirci_yilmaz post hoc Wilcoxon (Holm within three model pairs): ChatGPT vs Gemini adjusted p = 1.4e-12; ChatGPT vs Claude adjusted p = 1.2e-10; Gemini vs Claude adjusted p = 8e-11.

flesch_reading_ease post hoc Wilcoxon (Holm within three model pairs): ChatGPT vs Gemini adjusted p = 7.5e-07; ChatGPT vs Claude adjusted p = 3e-11; Gemini vs Claude adjusted p = 1.4e-12.

flesch_kincaid_grade post hoc Wilcoxon (Holm within three model pairs): ChatGPT vs Gemini adjusted p = 5.1e-08; ChatGPT vs Claude adjusted p = 0.488; Gemini vs Claude adjusted p = 1.4e-12.

Table S7. Reliability by presentation booklet

Booklet	n	QAMAI ICC(A,1)	Mean absolute rater diff.	Exact total agreement
1	42	0.642	1.286	13
2	42	0.907	0.500	27
3	42	0.593	1.095	15
4	42	0.687	0.786	19
5	42	0.769	0.810	20
6	42	0.729	1.143	9

Additional methodological notes

- Test–retest reliability was not assessed. Only the original two-rater scores for the 252 main responses enter these analyses.
- Question origin mapping: Q16, Q17, Q19 and Q21 were classified as direct-source questions in both languages in the locked question pool (Havuz_42 worksheet). These are all in domain A3. The restricted analysis is descriptive because n = 4 questions.
- The six de-identified rating booklets contain all 252 responses. The raw interface-to-blinded-text conversion may affect reference presentation, especially when citation chips were removed (see Online Resource 3 and the main manuscript limitations).

Table S8. Locked-pool direct-source subset: descriptive Turkish–English QAMAI differences

Question ID	TR mean	EN mean	TR – EN	Original source IDs
Q16	25.67	26.33	-0.67	S06; S07
Q17	24.33	24.83	-0.50	S06; S07
Q19	24.50	25.50	-1.00	S05; S06; S07
Q21	24.33	25.50	-1.17	S06; S07
Four-question subset	24.71	25.54	-0.83	S05, S06, S07

Values represent the mean across three chatbots after averaging the two raters for each response. Each question contributes three responses per language (12 per language in total). The restricted comparison was reported descriptively; no p value or equivalence decision was calculated for this subset. The word “direct” reflects the original locked-pool origin label; source questions included contextual additions specifying robotic surgery.