Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# 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

**zınar serhanoğlu**

dr.zinarserhanoglu@gmail.com

Hisar Intercontinental Hospital

**sema karakaş**

Hisar Intercontinental Hospital

**burcu akay serhanoğlu**

Haseki Eğitim ve Araştırma Hastanesi

**fulya gökdağlı sağır**

İstanbul Galata Üniversitesi

**özgür aslan**

Bağcılar Eğitim ve Araştırma Hastanesi

**basri çakıroğlu**

Hisar Intercontinental Hospital

**Research Article**

**Keywords:** large language models, robotic surgery, gynecology, patient education, guideline concordance, readability

**Posted Date:** November 26th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11749929/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Abstract

Background: How well chatbots answer surgical questions outside English is unclear. We compared the quality, guideline concordance and readability of Turkish and English answers to patient questions about robot-assisted gynecologic surgery. Methods: In this preregistered cross-sectional study, we submitted 42 benign and oncologic questions in each language to the free versions of ChatGPT, Gemini and Claude (252 responses). Two chatbot-blinded clinicians independently scored each response with the Quality Analysis of Medical Artificial Intelligence (QAMAI) tool (range 6–30) and rated potential harm, guideline concordance and adapted JAMA criteria. Equivalence was prespecified as ±1.5 points. Results: Mean QAMAI was 24.72 in Turkish and 25.66 in English (difference −0.94; 95% confidence interval [CI] −1.28 to −0.61; $p < 0.001$). The 90% CI (−1.22 to −0.67) lay within the equivalence margin, but the language effect differed by chatbot (interaction $p < 0.001$): +0.15 for ChatGPT, −1.51 for Gemini and −1.48 for Claude. The gap came mainly from the Sources (−0.71) and Completeness (−0.37) items. Separately rated accuracy did not differ, but rater agreement on it was poor (intraclass correlation 0.433). Both raters flagged one potentially harmful and two guideline-discordant responses, all from Gemini in Turkish. No English response met a sixth-grade reading level. Single-rater QAMAI reliability (0.728) was below the prespecified 0.75. Conclusions: The Turkish–English gap depended on the chatbot and reflected source reporting and completeness rather than accuracy, although accuracy was rated unreliably. Pooled equivalence does not make chatbots interchangeable across languages; source transparency, readability and occasional guideline-discordant answers remain concerns.


# Introduction

Robot-assisted surgery is used for hysterectomy, myomectomy, sacrocolpopexy and gynecologic oncology [1]. Superiority over conventional laparoscopy remains uncertain: a Cochrane review found similar complication rates (low-certainty evidence) but longer operating times [1], while robotic hysterectomy cost more [2]. In a US cohort of 533,258 elective benign hysterectomies, robotic use rose from 31.8% (2013) to 36.5% (2020) [3]. ACOG advises route selection by indication and patient factors rather than routine robotics for low-complexity cases [4]. Hospital websites have often promoted robotics without balanced risk disclosure [5].

The appropriateness of minimally invasive surgery differs by cancer type. In selected early endometrial cancers, minimally invasive surgery is guideline-supported [6]. In early cervical cancer requiring radical hysterectomy, LACC found worse survival with minimally invasive than open surgery [7, 8]; laparotomy remains the recommended route [9]. Selected low-risk early cervical cancers may instead be treated with simple hysterectomy, as defined by the 2026 ESMO guideline [9]. Whether robot-assisted radical hysterectomy provides comparable oncologic outcomes is being tested in RACC [10]. A prespecified secondary analysis found fewer mainly low-grade intraoperative complications but similar postoperative morbidity; its oncologic endpoint remains pending [11]. Evidence for minimally invasive interval cytoreduction in advanced ovarian cancer is also evolving [12]. Such benefits should not be generalized across indications.

Large language models (LLMs) are increasingly consulted for health information, despite public concerns about accuracy [13]. Recent studies of robotic urologic and colorectal procedures found variable response quality and source transparency, often with demanding reading levels [14–17]. Studies in gynecology and gynecologic oncology have reported heterogeneous rankings among chatbots and predominantly examined clinical scenarios or postoperative questions rather than patient decisions about operative routes [18–21].

Cross-lingual analyses report differences in the accuracy and verifiability of health-related LLM answers [22]. In Turkish–English comparisons, ChatGPT gave fewer correct Turkish answers to questions from pregnant women with rheumatic diseases [23], and Gemini and DeepSeek, but not ChatGPT, were less accurate in Turkish on common pregnancy questions [24]. Most patient resources on robotic gynecologic surgery are in English, and we found few Turkish society materials while developing our questions.

To our knowledge, no study has compared Turkish and English chatbot answers to patient questions across benign and oncologic indications in robotic gynecology. We therefore compared ChatGPT, Gemini and Claude on 42 patient questions in both languages. Prespecified primary analyses assessed the paired difference in Quality Analysis of Medical Artificial Intelligence (QAMAI) scores [25] and equivalence within ± 1.5 points. Secondary outcomes included factual accuracy, potential harm, guideline concordance, adapted JAMA transparency criteria [26] and readability.

## Materials and methods

# Study design and registration

In this cross-sectional study, each of 42 questions was submitted once in Turkish and once in English to three chatbots (ChatGPT, Gemini and Claude), giving 252 responses, and two clinicians rated every response independently. We submitted the protocol to OSF on 5 November 2026; the embargoed registration was approved on 8 November 2026 (view-only links for peer review; registration: https://osf.io/37eyb/overview?view_only=23479de649bd4f0f9bb79ca67bd9e5cc; project: https://osf.io/jgrnk/overview?view_only=c0d4c2b02e694b8794f89d74976594b0). We registered no amendments afterwards; deviations from the protocol are listed in Table 4. Reporting follows CHART [27] and the applicable STROBE items [28]. The study involved no patients or patient records.

The four preregistered hypotheses were: H1, a Turkish–English difference in QAMAI; H2, equivalence within ± 1.5 points; H3, more benefit overstatement and fewer alternative-route mentions in Turkish; and H4, chatbot differences in QAMAI and readability within each language.

## Question development

We drew the questions from patient materials published by professional societies, mainly ACOG, ESHRE, IUGA and ESMO. We also screened Turkish sources but found little relevant society material. Commercial, hospital and clinician-generated sources were excluded. Source links and selection rules appear in Online Resource 1.

Six domains contained seven questions each: benign hysterectomy (A1); myomectomy/endometriosis (A2); prolapse/sacrocolpopexy (A3); endometrial cancer (A4); cervical cancer (A5); and ovarian cancer/risk-reduction (A6). Two investigators independently screened 59 candidate questions (agreement 58/59; Cohen $\kappa = 0.96$) and resolved the single disagreement by consensus. None of the suitable candidates had surgeon experience, center or cost as its main topic, so we used the prespecified replacement questions. The 42-item pool was locked on 5 November 2026.

Two bilingual physicians aligned the Turkish and English versions by forward–backward translation and resolved discrepancies by consensus. In the locked pool, English items were designated direct (n = 5), adapted (n = 36) or translated (n = 1); Turkish items were direct (n = 4), translated (n = 36) or adapted (n = 2). Four items were classified direct in both languages, although contextual phrasing sometimes differed from source wording.

Fifteen questions (Q03, Q04, Q09, Q11, Q16, Q29–Q34, Q36, Q37, Q39 and Q42) were classified before querying as guideline-sensitive because correct answers needed to address when a robotic/minimally invasive route is not recommended, not standard or not demonstrably superior to laparoscopy.

## Chatbots and data collection

The free consumer web interfaces of ChatGPT (displayed label: GPT-6), Gemini (displayed label: Gemini 3, automatic model selection), and Claude (displayed label: Sonnet 5.5, medium) were queried using default settings. The underlying Gemini backend version was not shown. One investigator in Istanbul submitted each question verbatim in a new temporary chat, without follow-ups, and logged the first complete response. For all 126 question–chatbot pairs the Turkish version was submitted first and the English version second; this order was fixed rather than randomized. We used signed-in accounts with the interface language set to English, temporary chats, and memory and personalization switched off. Web search was left enabled in all three interfaces, but we did not record whether a search was actually run for a given response. Data were collected on 8–10 November 2026, and no response had to be requested again. One displayed-model entry was corrected during quality control; response text was unchanged.

## Pilot and rater calibration

Before the main study we generated 30 pilot answers to five questions outside the final pool. Both raters independently scored ten of these, selected at random (seed 20261105), for calibration. QAMAI single-

measure absolute-agreement ICC was 0.98 (95% CI 0.92–0.99), above the prespecified 0.75 threshold; no pilot answer entered the main analysis.

## De-identification, randomization and rating

Before scoring, a script removed identifying model references and interface artifacts and standardized formatting. Hyperlinks were removed from all responses, but visible source names and titles were kept as plain text. This affected citation chips in 61 ChatGPT responses (35 Turkish, 26 English), inline source links in 19 ChatGPT responses (9 Turkish, 10 English) and source lists in 14 Claude responses (9 Turkish, 5 English); no Gemini response contained a citation element. The raters could therefore see which sources were named but could not follow the links, which may have affected source-attribution scores. We extended the de-identification rules before rating began, without filing an OSF amendment (Online Resource 3).

Responses were randomized (seed 20261105), coded R001–R252 and divided into six booklets of 42. Two bilingual obstetrician–gynecologists rated all responses independently on 10–13 November 2026; they did not know which chatbot had produced a response but could see its language. The coordinator kept the key and took no part in scoring. Ratings were entered in validated spreadsheets rather than the preregistered online form (Table 4).

# Outcome measures

The primary outcome was QAMAI total (six 1–5 items assessing accuracy, clarity, relevance, completeness, sources and usefulness; range 6–30) [25]. Preregistered QAMAI grades were poor (6–11), fair (12–17), good (18–23), very good (24–29) and excellent (30). Since no response was poor or excellent, results are grouped as ≥ 24, 18 to < 24 and < 18. Secondary outcomes were an independent 1–5 accuracy rating, potential harm, B1 (mention of non-robotic or non-surgical alternatives), B2 (unqualified superiority over laparoscopy), B3 (guideline discordance), adapted four-item JAMA transparency criteria [26], and readability. B4 (surgeon or centre expertise) was exploratory.

Adapted JAMA criteria scored the presence of authorship, locatable attribution, disclosure of response limitations and currency (one point each). Rating anchors and reference guidance, including ACOG, ESGO, LACC and ESMO [4, 6, 7, 9], are in Online Resource 2.

Continuous scores were averaged across raters. Binary outcomes were counted as positive only if both raters scored yes (prespecified primary rule). Because this rule favors specificity and can undercount safety signals, we also report results under the at-least-one-rater rule.

Language-specific readability was calculated after prespecified text cleaning: Turkish Ateşman and Bezirci–Yılmaz [29, 30], and English Flesch Reading Ease and Flesch–Kincaid Grade Level [31, 32]. Scores were not compared across languages because the formulas differ. Response length was exploratory; because Turkish is agglutinative, word counts are not comparable across languages.

## Use of artificial intelligence tools

Claude (Anthropic) and ChatGPT (OpenAI) were used for syntax assistance in writing the analysis scripts in [R/SPSS/Python]. All code was reviewed, tested, and executed by the authors, who performed and verified all statistical analyses.

The authors used Claude (Anthropic) and ChatGPT (OpenAI) for language editing. The authors reviewed and edited all content and take full responsibility for the integrity and accuracy of the manuscript.

# Statistical analysis

Reliability was assessed using two-way random-effects, absolute-agreement single/average ICCs (95% CIs) for QAMAI, accuracy and JAMA totals, with ICC(A,1) ≥ 0.75 prespecified as acceptable. Weighted κ assessed QAMAI items, Cohen κ assessed binary items and Cronbach α assessed QAMAI internal consistency.

For H1, Turkish and English QAMAI totals were averaged across chatbots, yielding 42 paired question observations. We checked the paired differences for normality with the Shapiro–Wilk test and compared them with a paired t test, or with a Wilcoxon signed-rank test if normality was rejected. H2 used two one-sided tests (TOST) with the preregistered bounds of ± 1.5 points and a 90% confidence interval; the bound is 6.25% of the 24-point QAMAI score range. Chatbot-specific paired Wilcoxon tests received Holm correction.

A supportive random-intercept mixed-effects model included language, chatbot and their interaction, with question as the grouping factor. REML and Wald tests were used for main effects; the interaction was assessed by likelihood-ratio comparison of ML fits.

Secondary language contrasts used the paired approach; binary language effects used McNemar tests on 126 question–chatbot pairs (exact when fewer than 25 discordances). Cochran Q and Friedman tests compared chatbots within language, followed where indicated by Holm-adjusted paired tests.

Preregistered sensitivity analyses included the four direct-source bilingual items (descriptive because n < 15), an alternative binary coding rule and restrictions by question origin. Exploratory analyses examined clinical domain, benign versus oncologic indications, guideline-sensitive items, response length and B4. Two analyses were added post hoc and are labeled as such: chatbot-specific equivalence tests, and the primary comparison repeated on a five-item QAMAI total excluding the Sources item.

Tests were two-sided (α = 0.05) except TOST, with Holm correction within relevant families. No responses were excluded. Under preregistered assumptions, 42 pairs provided approximately 89% power for dz = 0.5 (two-sided α = 0.05) and 96% for equivalence (SD 2.55, true difference zero, margin ± 1.5). Reliability confidence intervals were estimated by 5,000 bootstrap resamples. Analyses were run in Python with NumPy 2.3.5, SciPy 1.17.0, statsmodels 0.14.6 and pandas 2.2.3; the analysis, readability

and randomization scripts are available in the OSF project. Additional procedures are described in Online Resource 3.

## Results

# Response characteristics

All 252 responses were analyzed (126 per language; 42 per chatbot–language combination). Median response lengths were 427 words in English and 323 in Turkish. Within English responses, ChatGPT produced the longest answers (median 493.5 words) and Gemini the shortest (378.0; Friedman $\chi^2$=18.77, $p < 0.001$). In Turkish, medians were 353.0, 302.5 and 341.0 words for ChatGPT, Gemini and Claude (Friedman $p = 0.203$). These comparisons were exploratory (Table 1).

# Inter-rater reliability

For QAMAI total, ICC(A,1) = 0.728 (95% bootstrap CI 0.639–0.795), below the prespecified 0.75 threshold, while ICC(A,2) = 0.842 (95% CI 0.780–0.886). Single-measure ICCs were 0.433 (95% CI 0.267–0.572) for separate accuracy and 0.727 (95% CI 0.647–0.790) for adapted JAMA total. Quadratically weighted κ ranged from 0.035 (relevance) to 0.844 (usefulness); the raters gave identical relevance scores to 176/252 responses (69.8%), and S.K. and B.A.S. gave the maximum score of 5 to 198 and 206 responses, respectively, so the low κ for relevance largely reflects a ceiling effect. Cronbach α was 0.357 for S.K. and 0.286 for B.A.S. Binary Cohen κ values were 0.665 (harm), 0.811 (B1), 1.000 (B2), 0.798 (B3) and 0.857 (B4). Absolute between-rater QAMAI disagreement did not correlate with booklet position ($\rho = 0.020$, $p = 0.758$; Online Resource 4, Table S1). Intra-rater (test–retest) reliability was not assessed.

# Overall quality

The overall mean rater-averaged QAMAI total was 25.19 (SD 1.72; median 25.5; range 16.5–29.0). English responses averaged 25.66 (SD 1.37) and Turkish responses 24.72 (SD 1.89). In total, 115/126 English responses and 99/126 Turkish responses scored ≥ 24; 11 and 26, respectively, scored 18 to < 24, and no English response and one Turkish response scored < 18 (Table 1). The mean adapted JAMA total was 1.08 of 4 points.

# Turkish versus English (H1 and H2)

Across the 42 question-level pairs averaged over chatbots, Turkish responses scored 0.94 points lower than English responses (Turkish minus English mean difference − 0.94; 95% CI − 1.28 to − 0.61; paired $t = -5.70$, $df = 41$, $p < 0.001$; Cohen $dz = -0.88$); the paired differences were compatible with normality (Shapiro–Wilk $p = 0.713$). The 90% CI (− 1.22 to − 0.67) lay entirely within the preregistered ± 1.5-point equivalence bounds (TOST $p < 0.001$). The pooled difference was therefore statistically significant and large in standardized terms, yet inside the prespecified margin.

Language differences varied by model: ChatGPT + 0.15 (Holm-adjusted $p = 0.328$), Gemini − 1.51 ($p < 0.001$) and Claude − 1.48 ($p < 0.001$; Wilcoxon; Fig. 1). Post hoc model-specific TOST showed

equivalence only for ChatGPT (90% CI − 0.203 to 0.512; p < 0.001), not for Gemini (− 2.053 to − 0.971; p = 0.515) or Claude (− 1.886 to − 1.067; p = 0.461). The mixed model estimated a mean language coefficient of − 0.94 (95% CI − 1.29 to − 0.60; p < 0.001) and a significant language-by-chatbot interaction (likelihood-ratio $\chi^2$=18.94, df = 2, p < 0.001).

Sources (QAMAI item 5; difference − 0.71, Holm-adjusted p < 0.001) and Completeness (item 4; −0.37, adjusted p < 0.001) contributed most to the pooled language gap. Together they sum to − 1.08; the total difference is smaller (− 0.94) because Clarity was slightly higher in Turkish (+ 0.14; adjusted p = 0.044). Accuracy, Relevance, Usefulness and separate accuracy (− 0.09; adjusted p = 0.333) were not significantly different after correction. Adapted JAMA differed by − 0.13 (unadjusted p = 0.077). In a post hoc paired test removing Sources from QAMAI, the five-item difference was − 0.238 (95% CI − 0.520 to 0.043; paired t p = 0.095), which shows that source reporting accounted for most of the gap (Online Resource 4, Table S2).

# Guideline-concordance outcomes and potential for harm (H3)

Under the prespecified both-rater rule, harmful content was flagged in 1 Turkish and 0 English responses, B2 overstatement in 1 and 0, and B3 discordance in 2 and 0 (Table 2). Under the at-least-one-rater rule, the corresponding Turkish/English counts were 1/1 (harm), 1/0 (B2), and 2/1 (B3). Alternative routes (B1) appeared in 109/126 (86.5%) Turkish and 104/126 (82.5%) English responses (McNemar p = 0.336); B2 did not differ (exact p = 1.000). H3 was not supported, although with a single positive event the overstatement comparison was uninformative.

All four flags fell on three Turkish Gemini responses. The answer to Q29 (cervical cancer; B3 and potential harm) presented perioperative advantages of robotic radical hysterectomy without discussing the LACC survival concern or the open-surgery standard. The answer to Q40 (ovarian cancer; B3) described robotic surgery as frequently preferred without distinguishing selected minimally invasive candidates from advanced-disease cytoreductive surgery. The answer to Q22 (endometrial cancer; B2) suggested benefits compared with conventional laparoscopy without qualifying the limited evidence for clinical superiority. Q40 and Q22 had not been classified as guideline-sensitive. Surgeon or center experience (B4, exploratory) was mentioned in 54/126 (42.9%) Turkish and 70/126 (55.6%) English responses (unadjusted McNemar p = 0.014).

# Differences between chatbots (H4)

In Turkish, QAMAI totals differed among the three chatbots (Friedman $\chi^2$ = 20.64, p < 0.001, Kendall W = 0.246). ChatGPT had the highest mean (25.52), followed by Claude (25.10) and Gemini (23.54). ChatGPT versus Gemini (Holm-adjusted p < 0.001) and Gemini versus Claude (p = 0.001) were significant; ChatGPT versus Claude was not (p = 0.064). In English (exploratory), QAMAI also differed among models (Friedman $\chi^2$ = 20.93, p < 0.001, Kendall W = 0.249), with Claude highest (26.57), followed by ChatGPT (25.37) and Gemini (25.05). Claude exceeded ChatGPT (adjusted p < 0.001) and Gemini (adjusted p <

0.001), while ChatGPT and Gemini did not differ (adjusted p = 0.308). Within-language between-model binary comparisons are provided in Online Resource 4, Table S4.

# Readability

In Turkish, median Ateşman scores were 59.30, 41.33 and 58.22 for ChatGPT, Gemini and Claude, respectively, and median Bezirci–Yılmaz scores were 10.05, 20.87 and 12.87 (Friedman p < 0.001 for both indices after correction across the two Turkish indices). In English, the corresponding median Flesch Reading Ease scores were 28.54, 17.94 and 40.42, and median Flesch–Kincaid grade levels were 11.44, 14.44 and 11.31 (adjusted p < 0.001 for both; Table 3). None of the 126 English responses had a Flesch–Kincaid grade level ≤ 6. Because the indices are language-specific, Turkish and English readability scores were not compared directly.

# Sensitivity and exploratory analyses

The four bilingual direct-source questions (Q16, Q17, Q19 and Q21; all pelvic-organ-prolapse items) had mean QAMAI scores of 24.71 in Turkish and 25.54 in English, a descriptive difference of − 0.83 (Online Resource 4, Table S8); no inferential test was performed. When each rater was analyzed separately, Turkish–English mean differences were − 1.03 (95% CI − 1.38 to − 0.69) for S.K. and − 0.86 (− 1.26 to − 0.46) for B.A.S. In exploratory domain-level comparisons (seven question pairs per domain), the language difference remained significant after Holm correction for A2 (− 1.17, p = 0.021) and A6 (− 1.38, p = 0.047). Guideline-sensitive questions scored 25.58 versus 24.97 for others (exploratory p = 0.019); benign and oncologic domain means were 25.12 and 25.26 (p = 0.840; Online Resource 4, Table S5).

Table 1
Response length, rater-averaged QAMAI scores and descriptive score bands by language and chatbot.

| Lang. | Chatbot | n | QAMAI mean ± SD | Words median | ≥ 24 | 18–<24 | < 18 |
|---|---|---|---|---|---|---|---|
| EN | ChatGPT | 42 | 25.37 ± 1.20 | 493.5 | 38 | 4 | 0 |
| EN | Gemini | 42 | 25.05 ± 1.19 | 378.0 | 36 | 6 | 0 |
| EN | Claude | 42 | 26.57 ± 1.25 | 434.5 | 41 | 1 | 0 |
| TR | ChatGPT | 42 | 25.52 ± 1.45 | 353.0 | 38 | 4 | 0 |
| TR | Gemini | 42 | 23.54 ± 2.14 | 302.5 | 23 | 18 | 1 |
| TR | Claude | 42 | 25.10 ± 1.42 | 341.0 | 38 | 4 | 0 |

Table 2
Potential for harm and guideline-related binary indicators by language and chatbot (both-rater rule: positive only if both raters scored yes).

| Lang. | Chatbot | Harm | B1 alternative | B2 overstatement | B3 discordance | B4 experience |
|---|---|---|---|---|---|---|
| EN | ChatGPT | 0 | 35 | 0 | 0 | 25 |
| EN | Gemini | 0 | 35 | 0 | 0 | 18 |
| EN | Claude | 0 | 34 | 0 | 0 | 27 |
| TR | ChatGPT | 0 | 35 | 0 | 0 | 17 |
| TR | Gemini | 1 | 34 | 1 | 2 | 13 |
| TR | Claude | 0 | 40 | 0 | 0 | 24 |

Table 3
Language-specific readability scores by chatbot, median (interquartile range).

| Lang. | Chatbot | Index 1 | Median (IQR) | Index 2 | Median (IQR) |
|---|---|---|---|---|---|
| TR | ChatGPT | Ateşman | 59.30 (56.65–64.10) | Bezirci–Yılmaz | 10.05 (8.83–11.41) |
| TR | Gemini | Ateşman | 41.33 (36.00–47.94) | Bezirci–Yılmaz | 20.87 (16.00–23.42) |
| TR | Claude | Ateşman | 58.22 (54.72–60.21) | Bezirci–Yılmaz | 12.87 (11.36–14.07) |
| EN | ChatGPT | Flesch Reading Ease | 28.54 (24.94–37.65) | Flesch–Kincaid grade | 11.44 (10.13–12.17) |
| EN | Gemini | Flesch Reading Ease | 17.94 (8.61–30.67) | Flesch–Kincaid grade | 14.44 (12.68–15.85) |
| EN | Claude | Flesch Reading Ease | 40.42 (35.47–47.57) | Flesch–Kincaid grade | 11.31 (10.14–12.12) |

Table 4
Deviations from the preregistered protocol and post hoc additions.

| Item | Preregistered plan | What was done | Handling and likely impact |
|---|---|---|---|
| Rating instrument | Online rating form | Validated spreadsheets | Procedural change; scoring items and response codes in both workbooks correspond to the study scoring guide. |
| De-identification rules | Rules as registered | Rules extended for newly observed artifacts before rating; no OSF amendment filed | Extension preceded rating; details in Online Resource 3 |
| Rating schedule | Scoring guide recommended an interval between booklets | All six booklets rated over four days (10–13 November 2026) | Rater fatigue possible; rater disagreement was not associated with booklet position (ρ = 0.020) |
| Inter-rater reliability | ICC(A,1) ≥ 0.75 prespecified as acceptable | Main-study ICC(A,1) = 0.728 | Averaging the two raters was prespecified; we report the shortfall and did not rescore. |
| Repeat rating | Not part of the registered protocol | No repeat rating performed | Not a protocol deviation; intra-rater (test–retest) reliability could not be estimated (see Limitations). |
| Model-label metadata QC | Not applicable | One model-label entry error corrected; response text unchanged | No effect on ratings |
| Additional analyses | Not registered | Chatbot-specific equivalence tests; five-item QAMAI total excluding Sources | Reported as post hoc |
| Rater blinding | Raters unaware of the assigned conditions during coding | Raters blinded to chatbot identity but not to response language | Language is evident from the text and could not be masked; expectation bias in the language comparison cannot be excluded (see Limitations) |

# Discussion

In this preregistered comparison, Turkish QAMAI scores were 0.94 points lower than English scores. This pooled difference fell within the ± 1.5-point margin, but the average hides a clear split between chatbots: ChatGPT showed no gap, whereas Gemini and Claude scored about 1.5 points lower in Turkish, right at the edge of the margin. The deficit came from the Sources and Completeness items; separately rated accuracy did not differ significantly, although the raters agreed poorly on that rating. Harm and guideline-

discordance flags were rare and confined to Turkish Gemini responses, and readability was poor in both languages.

The pooled equivalence result warrants caution: the ± 1.5-point margin was 6.25% of the scale range and 0.87 of the response-level SD. Pooled mean equivalence does not establish equivalence for individual models or responses.

Turkish answers scored lower on Sources and Completeness, while separately rated accuracy showed no detectable difference. Limited Turkish materials and differences in source handling are possible explanations, not demonstrated causes [22, 23]. We also stripped hyperlinks from all responses during de-identification, so the raters saw source names but could not check them; this may have affected Sources scores, although the procedure was the same for every chatbot and both languages. Adapted JAMA transparency was low overall (mean 1.08/4), but authorship and currency criteria developed for websites may show floor effects when applied to individual chatbot responses.

Earlier Turkish–English studies reported lower Turkish accuracy for ChatGPT [23], or for Gemini and DeepSeek but not ChatGPT [24]. Robotic urology and colorectal studies found variable transparency and demanding readability [14–17]. Gynecologic LLM studies chiefly examined clinical scenarios or postoperative advice [18–21]. Our study adds bilingual surgical-route questions and preregistered equivalence testing, although chatbot rankings are not directly comparable across studies.

How these findings should be read clinically depends on the indication. Minimally invasive surgery is standard for selected early endometrial cancers [6], whereas open radical hysterectomy remains the reference for early cervical cancer after LACC [7–9], and the role of minimally invasive interval cytoreduction in ovarian cancer is uncertain [12]. The flagged cervical cancer answer shows how a response can mislead without stating anything false. Perioperative advantages of robotic radical hysterectomy have some support from the RACC secondary analysis, which found fewer intraoperative complications [11], but the oncologic endpoint is pending and the answer omitted both the LACC survival signal and the open-surgery standard. Such omissions can matter more than factual errors and are poorly captured by a global quality score: guideline-sensitive questions scored slightly higher on QAMAI in exploratory analysis, and two of the three flagged answers were to questions not classified as guideline-sensitive. The small number of flags under the both-rater rule should therefore not be taken as a full count of consequential omissions.

No English response met the sixth-grade Flesch–Kincaid threshold. On the Ateşman scale, median Turkish responses from ChatGPT and Claude were moderately difficult (59.30 and 58.22), whereas Gemini's were difficult (41.33). Bezirci–Yılmaz medians suggested high-school level for ChatGPT (10.05), undergraduate level for Claude (12.87), and academic-level complexity for Gemini (20.87) [29, 30]. These labels come from formulas and do not tell us how well patients would actually understand the answers.

Strengths include preregistration, society-derived bilingual questions, blinded independent raters, prespecified equivalence testing and prospective identification of guideline-sensitive questions;

deviations appear in Table 4.

The study has several limitations. Each question was asked only once per chatbot and language, always in Turkish first, and we did not record whether a web search was run for a given response. QAMAI single-rater reliability fell below threshold; independent accuracy agreement and internal consistency were low, and test–retest reliability was unavailable. The absence of a detectable accuracy difference is therefore not evidence of equal accuracy, and the QAMAI total is better read as a composite index than as a single scale. Agreement was also much lower than in the ten-response calibration pilot, which may not have covered the range of content in the full sample. Both raters were co-authors, knew the answer language, and scored all booklets within four days. Gemini's underlying model was unspecified. Only four questions came directly from bilingual sources, all in one domain, and two of the flagged answers were to questions we had not classified as guideline-sensitive. We did not test patient comprehension, and the both-rater rule and the removal of hyperlinks may have led us to underestimate safety and attribution problems.

# Conclusions

Turkish answers scored lower than English answers overall, with model-dependent gaps: none detected for ChatGPT and approximately 1.5 points for Gemini and Claude. Differences predominantly concerned source reporting and completeness, not separately rated accuracy, although accuracy agreement was low. Pooled equivalence does not establish cross-language interchangeability. Rare harm and guideline-discordance flags involved oncologic questions where surgical route affects outcomes. Clinicians should verify chatbot advice against indication-specific evidence. Replication with repeated queries and more reliable ratings is needed.

# Declarations

**Funding.** The authors did not receive support from any organization for the submitted work.

**Competing interests.** The authors declare that they have no relevant financial or non-financial interests to disclose.

**Ethics approval.** Not applicable. The study evaluated publicly available chatbot outputs to predefined questions and involved no human participants, patient records or personal data.

**Consent to participate / consent for publication.** Not applicable.

Data availability. The locked question pool, de-identified chatbot responses and clinician ratings supporting this study are available from the corresponding author upon reasonable request. The analysis scripts are available in the OSF project (https://osf.io/jgrnk/overview?view_only=c0d4c2b02e694b8794f89d74976594b0).

Preregistration. The protocol was submitted to OSF on 5 November 2026 and the embargoed registration approved on 8 November 2026 (view-only links for peer review; registration: https://osf.io/37eyb/overview?view_only=23479de649bd4f0f9bb79ca67bd9e5cc; project: https://osf.io/jgrnk/overview?view_only=c0d4c2b02e694b8794f89d74976594b0). No subsequent protocol amendment was registered.

**Use of artificial intelligence.** The use of AI tools is described in the Methods. No AI tool is listed as an author, and the authors are accountable for the content of the manuscript.

# Author Contribution

Z.S. and B.Ç. conceived the study and participated in question screening and classification. Z.S. designed the methodology, collected the chatbot responses, curated the data, performed and verified the statistical analyses, coordinated the study, and wrote the original manuscript draft. S.K. and B.A.S. independently evaluated and scored the chatbot responses. F.G.S. and Ö.A. performed the forward–backward translation of the question set. B.Ç. supervised the study. All authors contributed to reviewing and editing the manuscript, read the final version, and approved its submission.

## References


1. Lawrie TA, Liu H, Lu D, Dowswell T, Song H, Wang L, Shi G (2019) Robot-assisted surgery in gynaecology. Cochrane Database Syst Rev 4(4):CD011422. https://doi.org/10.1002/14651858.CD011422.pub2
2. Wright JD, Ananth CV, Lewin SN et al (2013) Robotically assisted vs laparoscopic hysterectomy among women with benign gynecologic disease. JAMA 309(7):689–698
3. Cucinella G, Mathur M, Liu Y, Li Y, Schivardi G, Chiantera V et al (2026) Trends and comparative study of robotic-assisted benign hysterectomy among morbidly obese patients in United States. J Robot Surg 20(1):428. https://doi.org/10.1007/s11701-025-02991-1
4. American College of Obstetricians and Gynecologists' Committee on Gynecologic Practice; Society of Gynecologic Surgeons (2020) Robot-assisted surgery for noncancerous gynecologic conditions: ACOG Committee Opinion, Number 810. Obstet Gynecol 136(3):e22–e30. https://doi.org/10.1097/AOG.0000000000004048
5. Jin LX, Ibrahim AM, Newman NA, Makarov DV, Pronovost PJ, Makary MA (2011) Robotic surgery claims on United States hospital websites. J Healthc Qual 33(6):48–52. https://doi.org/10.1111/j.1945-1474.2011.00148.x
6. Concin N, Matias-Guiu X, Cibula D, Colombo N, Creutzberg CL, Ledermann J et al (2025) ESGO–ESTRO–ESP guidelines for the management of patients with endometrial carcinoma: update 2025. Lancet Oncol 26(8):e423–e435. https://doi.org/10.1016/S1470-2045(25)00167-6

7. Ramirez PT, Frumovitz M, Pareja R et al (2018) Minimally invasive versus abdominal radical hysterectomy for cervical cancer. N Engl J Med 379(20):1895–1904. https://doi.org/10.1056/NEJMoa1806395
8. Ramirez PT, Robledo KP, Frumovitz M, Pareja R, Ribeiro R, Lopez A et al (2024) LACC trial: final analysis on overall survival comparing open versus minimally invasive radical hysterectomy for early-stage cervical cancer. J Clin Oncol 42(23):2741–2746. https://doi.org/10.1200/JCO.23.02335
9. Lorusso D, Cibula D, Colombo I et al (2026) Cervical cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. ESMO Open 11(10):108364. https://doi.org/10.1016/j.esmoop.2026.108364
10. Karolinska Institutet Robot-assisted Approach to Cervical Cancer (RACC). ClinicalTrials.gov identifier: NCT03719547 [Internet]. [cited 2026 Nov 8]. Available from: https://clinicaltrials.gov/study/NCT03719547
11. Salehi S, Norberg Hardie A, Smolders RGV et al (2026) Perioperative outcomes and recovery after robot-assisted vs open radical hysterectomy: a secondary analysis of the RACC randomized clinical trial. JAMA Surg. Published online 7 October https://doi.org/10.1001/jamasurg.2026.4625
12. Rauh-Hain JA, Melamed A, Pareja R et al (2024) Laparoscopic cytoreduction after neoadjuvant chemotherapy in high-grade epithelial ovarian cancer: a LANCE randomized clinical trial. JAMA Netw Open 7(11):e2446325. https://doi.org/10.1001/jamanetworkopen.2024.46325
13. Presiado M, Montero A, Lopes L, Hamel L (2024) KFF Health Misinformation Tracking Poll: artificial intelligence and health information [Internet]. San Francisco: KFF; Aug 15 [cited 2026 Nov 8]. Available from: https://www.kff.org/health-information-trust/poll-finding/kff-health-misinformation-tracking-poll-artificial-intelligence-and-health-information/
14. Liu Y, Li R, Yu P, Zhang Y, Zhao A, Xiao X et al (2026) Reliability and readability of AI chatbot responses to patient questions about robot-assisted radical cystectomy. J Robot Surg 20(1):737. https://doi.org/10.1007/s11701-026-03698-7
15. Liu Y, Li R, Yu P et al (2026) Comparative performance of artificial intelligence chatbots in patient education for robot-assisted radical prostatectomy: quality, transparency and readability. J Robot Surg 20(1):857. https://doi.org/10.1007/s11701-026-03799-3
16. Wang L, Liu H, Yang J, Jiao P, Chen S, Si X et al (2026) Guideline-based evaluation of large language models for patient counselling before robot-assisted radical prostatectomy: accuracy, safety, and readability. J Robot Surg 20(1):888. https://doi.org/10.1007/s11701-026-03859-8
17. Yu J, Li M, Qin XZ, Zhou DB (2026) Quality, transparency, and readability of AI chatbot responses for patient counseling in robot-assisted colorectal cancer resection: a cross-sectional evaluation. J Robot Surg 20(1):896. https://doi.org/10.1007/s11701-026-03874-9
18. Voigt PC, Sharma RS, Chaudhari A, Tsai S, Milad MP, Yang LC (2026) Evaluation of ChatGPT and Gemini in answering patient questions after gynecologic surgery. Appl Clin Inf 17(2):172–176. https://doi.org/10.1055/a-2818-1611

19. Anastasio MK, Peters P, Foote J, Melamed A, Modesitt SC, Musa F et al (2024) The doc versus the bot: a pilot study to assess the quality and accuracy of physician and chatbot responses to clinical questions in gynecologic oncology. Gynecol Oncol Rep 55:101477
20. Gumilar KE, Indraprasta BR, Faridzi AS, Wibowo BM, Herlambang A, Rahestyningtyas E et al (2024) Assessment of large language models (LLMs) in decision-making support for gynecologic oncology. Comput Struct Biotechnol J 23:4019–4026. https://doi.org/10.1016/j.csbj.2024.10.050
21. Kaplan ST (2025) Artificial intelligence and gynecological oncology: a comparative study of ChatGPT Omni and Gemini Pro across repeated intervals with case-scenario and open-ended queries. Oncol Res Treat 48(6):325–331. https://doi.org/10.1159/000545231
22. Jin Y, Chandra M, Verma G, Hu Y, De Choudhury M, Kumar S (2024) Better to ask in English: cross-lingual evaluation of large language models for healthcare queries. In: Proceedings of the ACM Web Conference 2024 (WWW '24). :2627–2638. https://doi.org/10.1145/3589334.3645643
23. Özdemir Ulusoy B, Ulusoy CO (2025) Evaluating ChatGPT's effectiveness in providing medical information for pregnant women with rheumatic diseases. Jinekoloji-Obstetrik ve Neonatoloji Tıp Dergisi 22(1):38–44. https://doi.org/10.38136/jgon.1581349
24. Tığlı A, Baykuş Y, Deniz R, Üstebay S, Bektaş MH, Çöllüoğlu Ç (2026) When artificial intelligence speaks for the obstetrician: multilingual accuracy on real patient questions. Hipokrat Tıp Dergisi 6(2):61–69. https://doi.org/10.58961/hmj.1916158
25. Vaira LA, Lechien JR, Abbate V, Allevi F, Audino G, Beltramini GA et al (2024) Validation of the Quality Analysis of Medical Artificial Intelligence (QAMAI) tool: a new tool to assess the quality of health information provided by AI platforms. Eur Arch Otorhinolaryngol 281(11):6123–6131. https://doi.org/10.1007/s00405-024-08710-0
26. Silberg WM, Lundberg GD, Musacchio RA (1997) Assessing, controlling, and assuring the quality of medical information on the Internet: caveant lector et viewor—let the reader and viewer beware. JAMA 277(15):1244–1245. https://doi.org/10.1001/jama.1997.03540390074039
27. CHART Collaborative (2025) Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. BMJ Med 4(1):e001632. https://doi.org/10.1136/bmjmed-2025-001632
28. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP (2007) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 370(9596):1453–1457
29. Ateşman E (1997) Türkçede okunabilirliğin ölçülmesi. Dil Dergisi 58:71–74
30. Bezirci B, Yılmaz AE (2010) Metinlerin okunabilirliğinin ölçülmesi üzerine bir yazılım kütüphanesi ve Türkçe için yeni bir okunabilirlik ölçütü. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik. Dergisi 12(3):49–62
31. Flesch R (1948) A new readability yardstick. J Appl Psychol 32(3):221–233
32. Kincaid JP, Fishburne RP, Rogers RL, Chissom BS (1975) Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted

personnel. Research Branch Report 8–75. Millington (TN): Naval Technical Training Command

# Figures

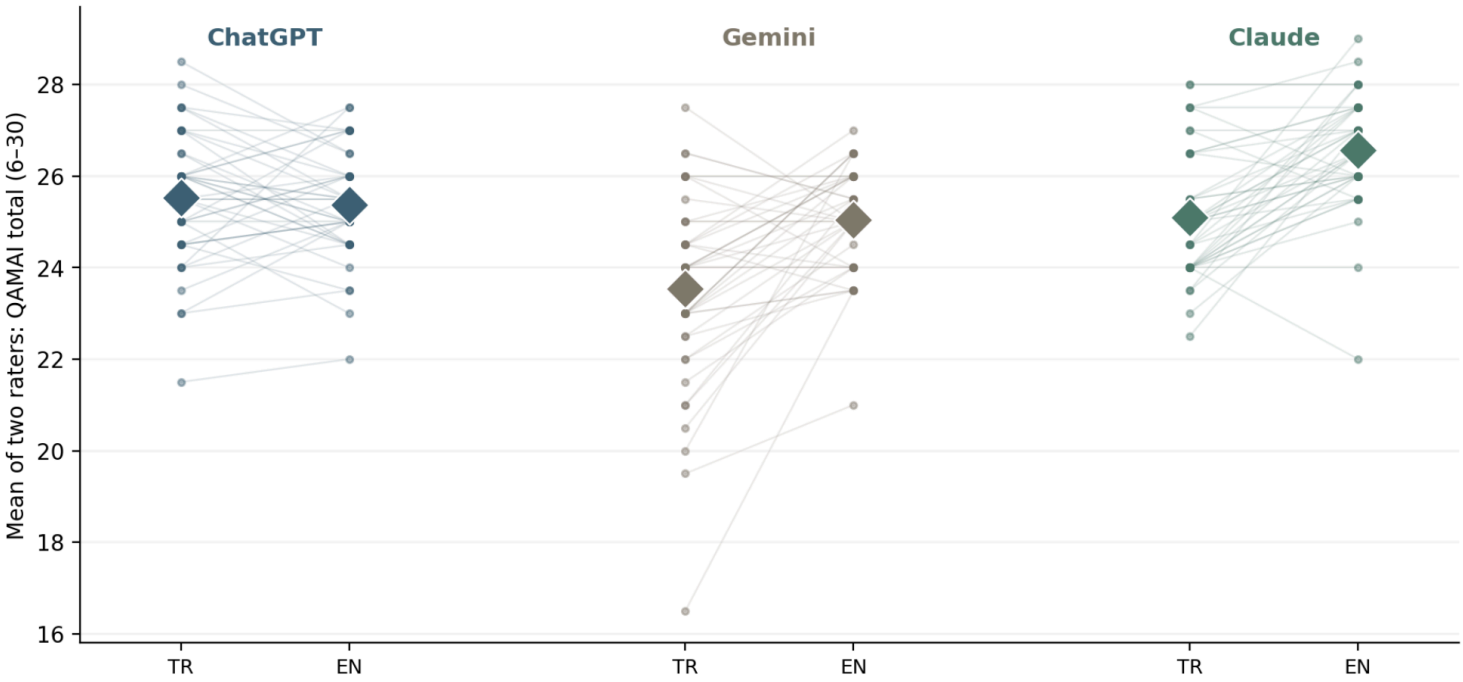


**Figure 1**

Paired question-level QAMAI scores (mean of two raters) by chatbot and language. Diamonds denote means; faint lines connect the Turkish (TR) and English (EN) responses to the same question.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1QuestionPool.pdf
- ESM2ScoringGuide.pdf
- ESM3ExtendedMethods.pdf
- ESM4StatisticalTables.pdf