Online Resource 3 — Extended study methods and de-identification audit

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

Extended methods accompanying the final manuscript. Procedural departures from preregistration are described as observed and were not retroactively registered.

Materials and methods

Study design and registration

This was a cross-sectional, comparative evaluation of chatbot outputs with a within-question design and two crossed factors: language (Turkish, English) and chatbot (ChatGPT, Gemini, Claude). Each of 42 patient-oriented questions was submitted once in each language to each chatbot, yielding 252 responses, and every response was rated by two independent raters. The protocol, hypotheses and analysis plan were submitted for preregistration to the Open Science Framework (OSF) on 5 November 2026; the embargoed registration was approved on 8 November 2026 (registration: https://osf.io/jgrnk; project: https://osf.io/37eyb). Initial submission preceded the first main-study chatbot query. No subsequent protocol amendment was registered. Procedural deviations observed during data collection and rating are described transparently below. The study is reported according to the Chatbot Assessment Reporting Tool (CHART) [27] and, where applicable, the STROBE recommendations [28]. The study involved no enrolled patients, clinical records or personal patient data; ethics committee approval was not sought for this evaluation of chatbot outputs.

Four hypotheses were preregistered. H1 (primary): QAMAI total scores differ between the Turkish and English versions of the same question (two-sided). H2: the mean Turkish–English difference in QAMAI total lies within an equivalence margin of ±1.5 points. H3: compared with English responses, Turkish responses more often overstate the benefits of robotic surgery and less often mention alternative surgical routes. H4: the three chatbots differ in QAMAI total and readability within each language.

Question development

Questions were derived exclusively from publicly available patient-education materials of professional societies and their patient organizations. The English-language materials searched were those of ACOG, the American Association of Gynecologic Laparoscopists, the European Society of Human Reproduction and Embryology (ESHRE), ESMO, ESGO and its patient network ENGAGe, the Society of Gynecologic Oncology and the Foundation for Women's Cancer, the Royal College of Obstetricians and Gynaecologists, the International Urogynecological Association

(IUGA) and the American Urogynecologic Society; the final pool drew on ACOG, ESHRE, IUGA and ESMO materials. Turkish-language sources searched were the Turkish Society of Obstetrics and Gynecology (TJOD) and the Turkish patient leaflets of IUGA. Apart from one IUGA leaflet, no Turkish society material addressing robotic gynecologic surgery was identified. When an original society website could not be reached, a verbatim copy hosted elsewhere was used and the original link was recorded. Hospital, clinic and manufacturer websites, patient forums and clinician-generated questions were excluded.

The pool was structured into six clinical domains with seven questions each: benign hysterectomy (A1), myomectomy and endometriosis surgery (A2), sacrocolpopexy for pelvic organ prolapse (A3), endometrial cancer including staging and sentinel lymph node mapping (A4), cervical cancer and radical hysterectomy (A5), and ovarian cancer and risk-reducing surgery (A6). Within each domain, six questions were to cover predefined strata: (1) definition and how the procedure works; (2) candidacy and indications; (3) comparison with alternative routes; (4) risks, complications and oncological safety; (5) surgeon experience, centre selection and cost; and (6) recovery, daily life and follow-up. The seventh question was the most prominent remaining patient question in that domain. Two evaluators (Z.S. and B.Ç.) independently screened 59 candidate questions against prespecified eligibility criteria: the question could plausibly be asked by a patient, could be answered without visual material, belonged to a single domain and was traceable to a society source through a public link. Agreement was 58/59 (Cohen's $\kappa = 0.96$), and the single disagreement was resolved by consensus. No candidate in stratum 5 was judged eligible, so, as specified in the protocol, these positions were filled with the most prominent remaining question in the same domain. The final pool comprised six questions on definition, eight each on candidacy, comparison with alternatives, risks and recovery, and four additional domain-specific questions. The pool was locked on 5 November 2026 and deposited with the registration (Supplementary File 1).

Turkish and English versions were prepared as content-equivalent items, and the origin of each version was recorded as direct (classified by the locked question pool as derived from a society-language question with minor contextual edits), translated, or adapted (formulated from society content that contained no corresponding question). The English version was direct for 5 questions, adapted for 36 and translated for 1; the Turkish version was direct for 4, translated for 36 and adapted for 2. Only four questions were direct in both languages. Translations were produced by two bilingual physicians (F.G.S. and Ö.A.) using forward–backward translation, and discrepancies were resolved by consensus; abbreviations were spelled out.

Before registration, each question was classified as guideline-sensitive or not by applying one criterion: whether a correct answer must state that the robotic or minimally invasive route is not recommended, not standard, or not shown to be superior to conventional laparoscopy. The two evaluators classified all 42 questions, and 15 formed the guideline-sensitive subset (Q03, Q04, Q09, Q11, Q16, Q29–Q34, Q36, Q37, Q39 and Q42).

Chatbots and data collection

Three publicly available chatbots were queried through signed-in free consumer web accounts with English-language interfaces and default settings: ChatGPT (OpenAI), Gemini (Google) and

Claude (Anthropic). Web-search functionality remained enabled but whether a specific answer used it was not logged. Temporary chats were used with memory and personalization disabled. Turkish was always queried before English for each question–chatbot pair (fixed order). The model name displayed in the interface was recorded for every query: "GPT-6" for ChatGPT, "Gemini 3" with automatic model selection for Gemini (the specific underlying model is not displayed in the free interface), and "Sonnet 5.5" (medium setting) for Claude. A single investigator (Z.S.) entered all queries from Istanbul, Türkiye. Each question was entered in a new chat in the chatbot's temporary (memory-free) mode, with memory and personalization switched off. Questions were submitted verbatim, without instructions, greetings or follow-up messages, and the first complete response was copied verbatim into a structured query log together with the date, time, displayed model name and confirmation of temporary mode. Queries were organized in seven blocks of 36, and the Turkish and English versions of each question were submitted to the same chatbot consecutively. All 252 responses were obtained between 8 November 2026 (20:50) and 10 November 2026 (13:00, local time), within the preregistered limit of seven consecutive days. No model switch, technical error or re-request occurred. One data-entry error in the displayed-model field was identified during quality control, confirmed against the investigator's record and corrected in the event log; the response text was not affected.

Pilot and rater calibration

On 5 November 2026, five patient questions outside the final pool were submitted to the three chatbots in both languages through the same workflow, producing 30 pilot responses. Ten of these were selected at random (seed 20261105), de-identified and scored independently by both raters using the written scoring guide. The single-measure, absolute-agreement intraclass correlation coefficient, ICC(A,1), for the QAMAI total was 0.98 (95% CI 0.92–0.99), exceeding the preregistered threshold of 0.75; main querying therefore began after a single calibration round. In this round, ICC(A,1) was 0.74 for independent accuracy and 1.00 for the JAMA total, and 148 of 160 item ratings were identical between raters. Pilot responses were not included in any analysis.

De-identification, randomization and rating

Before rating, responses were processed with a Python script that replaced model and company names and self-references with a neutral placeholder, removed interface artefacts, and standardized formatting (headings, emphasis, tables and list markers converted to plain text), without changing the wording; every change was logged per response. During blinding of the main-study responses, interface artefacts not seen in the pilot were identified, mainly inline image-carousel links and counters, citation-chip lines, pipe-separated source cards, form buttons, markdown escape characters and LaTeX notation. The rules were extended to cover these before any rating took place, and these changes were applied before rating; they were not submitted as an OSF registration amendment. Beyond the original rules, only the captions of images that were not displayed to raters were removed. An automated search of the blinded texts for model, company and product names returned no matches. In the coordinator log, citation-chip elements were recorded as removed from 61 ChatGPT responses (35 Turkish, 26 English). The effect of these removals on QAMAI Sources and JAMA attribution cannot be reconstructed from blinded texts or rater scores, and no rescoring was performed.

Each response received a unique code (R001–R252) generated before data collection with a Python script (seed 20261105). The randomization key was held only by the study coordinator (Z.S.), who did not rate any response. Responses were presented in code order in six booklets of 42, so that chatbots, languages and domains were intermixed. The two raters were bilingual specialists in obstetrics and gynecology with at least five years of clinical experience: S.K., a gynecologic oncologist, and B.A.S., an obstetrician-gynecologist with experience in robotic gynecologic surgery as both bedside assistant and console surgeon. Neither took part in question development, querying or de-identification, and neither had access to the key. Raters were blinded to the chatbot but, necessarily, not to the language of the response. They scored independently, without consulting chatbots, in a spreadsheet with drop-down validation for every item and locked code and total columns. Use of a spreadsheet rather than the structured online form described in the preregistration was a procedural deviation. The two evaluators completed the ratings between 10 and 13 November 2026; individual session times were not documented in the supplied workbooks.

Outcome measures

The primary outcome was the total score of the Quality Analysis of Medical Artificial Intelligence (QAMAI) tool [25], which rates accuracy, clarity, relevance, completeness, sources and usefulness on 1–5 scales (total 6–30). Secondary outcomes were an independent rating of factual accuracy (1, completely incorrect, to 5, completely correct); potential for harm (yes or no: content that could lead to clinically meaningful or serious harm if a patient acted on it); three guideline-concordance items, each rated yes or no: B1, the response names at least one non-robotic route or a relevant non-surgical option; B2, the response states or implies better clinical outcomes than conventional laparoscopy without stating that comparative evidence is limited or shows similar outcomes; and B3, the response contradicts a reference document listed in the scoring guide; the JAMA benchmark criteria [26] adapted to chatbot output; and readability. An exploratory item (B4) recorded whether the response advised the patient to consider the surgeon's or centre's experience, specialization or case volume, or stated that outcomes depend on it.

The four JAMA criteria were adapted as follows, each scored 1 if met: authorship, when the response names the organization or author responsible for a specific statement; attribution, when it gives at least one locatable reference (a named document, study, dated guideline or link); disclosure, when it states its own nature or limits (for example, that it is general information and not a substitute for medical advice); and currency, when it dates the information it presents. The JAMA criteria were scored on presence only; correctness was captured by the QAMAI and accuracy ratings. Raters judged responses against their clinical knowledge and the reference documents listed in the scoring guide, which included ACOG Committee Opinion 810 [4], the ESGO–ESTRO–ESP endometrial carcinoma guidelines [6], the LACC trial [7] and the ESMO cervical cancer guideline [9]. The full scoring guide, with anchors and examples, is provided in Online Resource 2.

For continuous outcomes, the mean of the two raters' scores was analyzed. Binary outcomes were coded "yes" only when both raters answered "yes" (conservative rule, primary); the liberal rule (at least one "yes") was used in a sensitivity analysis.

Readability was calculated from the de-identified texts with a Python script implementing the preregistered preprocessing rules: markdown and emoji removed, link text kept and addresses removed, terminal punctuation and line breaks treated as sentence boundaries, and number-only tokens excluded. Turkish responses were scored with the Ateşman Reading Ease formula [29] and the Bezirci–Yılmaz readability score [30], counting syllables as vowels. English responses were scored with the Flesch Reading Ease [31] and Flesch–Kincaid Grade Level [32], with syllables counted using textstat 0.7.13. Because the formulas are language-specific, Turkish and English readability scores were not compared with each other. Response length in words was recorded as an exploratory variable.

Statistical analysis

Inter-rater reliability was assessed before aggregation using two-way random-effects, absolute-agreement ICCs, single and average measures with 95% CIs, for the QAMAI total, independent accuracy and the JAMA total; a single-measure ICC of at least 0.75 was considered acceptable. Quadratically weighted kappa was calculated for each QAMAI item, Cohen's kappa and percent agreement for each binary item, and Cronbach's alpha of the QAMAI items for each rater.

For the primary analysis (H1), rater-averaged QAMAI totals were averaged across the three chatbots, giving one Turkish and one English score per question (42 pairs). The distribution of the paired differences was examined with the Shapiro–Wilk test; normally distributed differences were compared with a paired t test (effect size Cohen's dz) and otherwise with the Wilcoxon signed-rank test (effect size rank-biserial correlation). Equivalence (H2) was tested on the same pairs with two one-sided tests (TOST) against bounds of −1.5 and +1.5 points, concluding equivalence when the 90% CI lay entirely within the bounds. A non-significant H1 result was to be interpreted as evidence of no difference only if equivalence was shown. The language effect within each chatbot was tested with Wilcoxon signed-rank tests (n = 42 each) with Holm correction across chatbots.

A supportive linear mixed-effects model was fitted to the 252 rater-averaged QAMAI totals, with language (coded −0.5 English, +0.5 Turkish), chatbot (centred contrasts) and their interaction as fixed effects and a random intercept for question. The model was estimated by restricted maximum likelihood; fixed effects were tested with Wald tests, and the language × chatbot interaction with a likelihood-ratio test comparing maximum-likelihood fits with and without the interaction.

Secondary comparisons between languages for the six QAMAI items, independent accuracy and the JAMA total followed the primary approach. Binary outcomes (harm, B1–B3) were compared between languages with McNemar's test on question–chatbot pairs (n = 126), using the exact test when discordant pairs numbered fewer than 25; H3 was tested on B1 and B2. Between-chatbot differences in binary outcomes within each language were tested with Cochran's Q (n = 42), followed, if significant, by pairwise McNemar tests with Holm correction. Between-chatbot differences in QAMAI total (H4) and readability within each language were tested with the Friedman test, followed, if significant, by pairwise Wilcoxon signed-rank tests with Holm correction.

Prespecified sensitivity analyses repeated the primary comparison in questions that were direct in both languages (to be analyzed inferentially only if at least 15 such questions existed; with four, it was reported descriptively), repeated all binary analyses under the liberal rule, and repeated the primary comparison after excluding questions whose Turkish version was translated or adapted. Exploratory analyses, labelled as such, compared languages within each domain, benign (A1–A3) with oncology (A4–A6) domains, the guideline-sensitive subset with the remaining questions (B2, B3 and QAMAI total), response length and B4 between languages and chatbots, and chatbots within the English arm.

All tests were two-sided with $\alpha = 0.05$, except the one-sided components of TOST. Holm correction was applied within prespecified families of secondary and exploratory comparisons (QAMAI items and accuracy; binary outcomes between languages; binary outcomes between chatbots; readability indices; domain-level comparisons), and both unadjusted and adjusted p values are reported. No responses were excluded; refusals or partial refusals would have been scored as responses. With 42 pairs, the estimated power of a two-sided paired t test for $dz=0.5$ and $\alpha=0.05$ is approximately 88.6%; the original design stated 80% power at 34 pairs and, assuming a standard deviation of paired differences of 2.55 points (the preregistered planning estimate) and a true difference of zero, approximately 96% power for the equivalence test. Main-study ICC 95% confidence intervals were derived from 5,000 response-level bootstrap resamples. Analyses of the supplied workbooks were implemented in Python 3.13.5 using SciPy 1.17.0, statsmodels 0.14.6 and NumPy 2.3.5. The analysis scripts are retained by the authors and are available from the corresponding author upon reasonable request.