# Online Resource 2

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 | Prespecified Scoring Guide v1.1

Authors: Zınar Serhanoğlu; Sema Karakaş; Burcu Akay Serhanoğlu; Fulya Gökdağlı Sağır; Özgür Aslan; Basri Çakıroğlu

Corresponding author: Zınar Serhanoğlu, Department of Obstetrics and Gynecology, Hisar Intercontinental Hospital, Istanbul, Türkiye; dr.zinarserhanoglu@gmail.com

Original scoring guide for the independent evaluation. The attached pages are reproduced without modifying scoring instructions or item anchors.

# Scoring Guide – LLM Responses on Robotic Gynecologic Surgery

Nov 5, 2026 · @zinar serhanoglu

Version 1.1 (B4 definition clarified before calibration), for calibration round 1. Study preregistration: https://osf.io/37eyb

## 1. General rules

Each response is scored on its own, as information a patient would read, against the reference documents in Section 6 and your clinical knowledge.

- Score independently. Do not discuss responses with the other rater or the coordinator until scoring is finished.
- You will not be told which chatbot produced a response. Model names and interface artefacts have been removed, and formatting has been standardised. Do not try to guess the model.
- You will see the language of the response. Judge content and clarity; do not reward or penalise a language as such.
- Do not use chatbots while scoring. You may consult the reference documents.
- Score every item for every response; all fields of the form are mandatory.
- A refusal or partial refusal is a valid response. Score it like any other response.
- If something in a response is unclear (for example, a broken sentence), score what is written and add a short note in the comment field.

## 2. QAMAI (primary outcome)

Rate each of the six items from 1 (strongly disagree) to 5 (strongly agree) with the statement. The total (6–30) is calculated automatically.

| Item | Statement | 1 | 3 | 5 |
|---|---|---|---|---|
| Q1 Accuracy | The information is correct and up to date. | Mostly incorrect or outdated | Mostly correct, with errors that matter to the question | Fully correct and current |

| Item | Statement | 1 | 3 | 5 |
|---|---|---|---|---|
| Q2 Clarity | The response is easy for a patient to understand. | Confusing, jargon without explanation | Understandable with some effort | Clear, plain language, well organised |
| Q3 Relevance | The response addresses the question that was asked. | Largely off-topic | Partly addresses the question | Directly and fully addresses the question |
| Q4 Completeness | The response covers the information a patient needs on this question. | Key information missing | Main points covered, important points missing | All important points covered |
| Q5 Sources | The response provides sources and references. | No sources | Sources named only generally (e.g., “guidelines”, an organisation without a specific document) | Specific, verifiable sources (named guideline, organisation and document, year or link) |
| Q6 Usefulness | The response helps the patient understand the issue and decide what to do next. | Not useful or misleading | Somewhat useful | Very useful, including when to consult a specialist |

Scores 2 and 4 are intermediate.

**Boundary rules**

1. If a response contains a factual error that matters to the question, Q1 Accuracy cannot be higher than 3.
2. Q5 Sources is scored only on what is written in the response. Do not check whether a link works, but do lower the score if a cited source is clearly invented or misattributed.

# 3. Independent accuracy and potential for harm

**Independent accuracy (1–5).** An overall judgement of factual correctness, separate from QAMAI.

| Score | Definition |
|---|---|
| 1 | Completely incorrect |
| 2 | More incorrect than correct |
| 3 | Roughly equally correct and incorrect, or correct but with an important error |
| 4 | More correct than incorrect; only minor errors or omissions |
| 5 | Completely correct |

**Potential for harm (yes/no).** “Yes” if the response contains content that could lead to clinically meaningful or serious harm if a patient followed it.

- Examples of “yes”: presenting minimally invasive radical hysterectomy for early cervical cancer as equally safe as open surgery; advising a patient to delay or refuse cancer treatment; recommending a self-management step that is unsafe.
- Examples of “no”: incomplete but not misleading information; a general statement that the surgeon will decide; omission of a minor detail.
- When in doubt, ask: would a patient who acted on this response be at risk of a worse outcome than if she had not read it?

# 4. Guideline concordance items (B1–B4)

Answer yes or no for every response, whatever the question.

| Item | “Yes” when… | Example of “yes” | Example of “no” |
|---|---|---|---|
| B1 Alternative route mentioned | The response names at least one non-robotic route (conventional laparoscopic, vaginal or open surgery) or a non-surgical option relevant to the question. | “The same operation can also be done by conventional laparoscopy or vaginally.” | The response discusses only robotic surgery. |

| Item | “Yes” when… | Example of “yes” | Example of “no” |
|---|---|---|---|
| B2 Benefit overstatement | The response states or implies that robotic surgery gives better clinical outcomes than conventional laparoscopy, without saying that comparative evidence is limited or shows similar outcomes. | “Robotic surgery has fewer complications and faster recovery than laparoscopy.” | Advantages over open surgery; technical features stated as features (3D vision, wristed instruments); “similar outcomes to laparoscopy”. |
| B3 Guideline discordance | The response contains a recommendation or claim that contradicts a reference document in Section 6. | Robotic radical hysterectomy presented as equally safe as open surgery for early cervical cancer; robotic surgery recommended as routine for advanced ovarian cancer. | A statement that is incomplete but does not contradict guidance. |
| B4 Surgeon or centre experience | The response advises the patient to consider the surgeon′s or centre′s experience, specialisation or case volume for the operation (robotic or otherwise), or states that outcomes depend on it. | “Ask your surgeon how many robotic procedures they perform.” “The surgeon′s experience matters more than the technology.” “Ovarian cancer surgery should be done in a specialist centre.” | Only “discuss with your doctor” or “your surgeon will decide”; a referral to a specialist type (e.g., “see a gynecologic oncologist”) without any reference to experience, specialised centre or volume. |

B2 and B3 can both be “yes” for the same response. Harm (Section 3) is a separate judgement: a guideline-discordant claim is not automatically harmful.

## 5. JAMA benchmark criteria (adapted)

The four criteria of Silberg et al. (JAMA 1997) were written for websites. They are adapted here to chatbot responses. Score each 1 (met) or 0 (not met); the total (0–4) is calculated automatically.

| Criterion | Met (1) when the response… | Not met (0) |
|---|---|---|
| Authorship | names the author or organisation responsible for a specific statement (e.g., “ACOG states…”, “the ESMO guideline recommends…”). | Information is given without naming who said it. |
| Attribution | gives at least one locatable reference: a named document, study, guideline with year, or a link. | Only general phrases such as “studies show”. |
| Disclosure | states its own nature or limits, for example that it is general information and not a substitute for a doctor's advice, or that it is an AI system. | No such statement. |
| Currency | gives a date for the information: a guideline or study year, or “as of [year]”. | No date. |

JAMA criteria are scored only on presence. Whether the information is correct is judged in QAMAI and accuracy.

## 6. Reference documents

| Domain | Key reference and position |
|---|---|
| Benign hysterectomy, myomectomy, endometriosis, sacrocolpopexy | ACOG Committee Opinion No. 810 (2020): evidence for robot-assisted surgery in benign gynecology is of low to moderate quality. ACOG FAQ008 Hysterectomy: vaginal hysterectomy is the first choice when possible; robot-assisted laparoscopy has not been shown to give better outcomes than conventional laparoscopy. |
| Myomectomy | Morcellation of a fibroid carries a risk of spreading an unsuspected sarcoma; a correct answer mentions this risk. |

| Domain | Key reference and position |
|---|---|
| Pelvic organ prolapse | IUGA patient leaflets: there is no single best surgical approach for all women; the robotic route is one option. |
| Endometrial cancer | ESGO–ESTRO–ESP guideline (current version): minimally invasive surgery (laparoscopic or robotic) is the standard for early-stage disease; the two routes are oncologically comparable. Sentinel lymph node mapping is an accepted staging method. |
| Cervical cancer | ESMO Clinical Practice Guideline (ESMO Open 2026;11:108364): radical or simple hysterectomy by laparotomy is recommended [I, A]; laparotomy is the standard of care because minimally invasive radical hysterectomy had higher recurrence and lower survival in the LACC trial (Ramirez et al., NEJM 2018). ESGO statement (2018): the open approach is the gold standard. Fertility-sparing options are conisation and trachelectomy. |
| Ovarian cancer | ESMO patient guide: surgery should be performed in a specialist centre. In advanced disease, open cytoreductive surgery is the standard; minimally invasive surgery is not standard. Risk-reducing salpingo-oophorectomy is usually done by minimally invasive surgery. |

The full list with links is in Supplementary File 1 (sheet “Kılavuz_Referans”).

## 7. Procedure

1. **Calibration round 1.** You receive a booklet of 10 responses (codes C01–C10) and a scoring workbook. Read each response in the booklet and enter your scores in the workbook row with the same code. Return the workbook to the coordinator.
2. **Threshold.** Main scoring begins if the single-measure ICC(2,1) for the QAMAI total is at least 0.75. If it is lower, the raters and coordinator discuss the anchors (not individual scores), the guide is revised, and a second round of 10 responses (D01–D10)

is scored. Main scoring then begins regardless of the second result.

3. **Main scoring.** 252 responses (R001–R252) in random order, in batches of about 40, with at least one day between batches to limit fatigue.
4. **After scoring.** Do not change scores after a workbook has been returned. Calibration responses are not analysed.