EURJIMR | European Journal of Innovative Medical Research

Type LETTER TO EDITOR
Published 05.01.2026
DOI 10.65495/eurjimr.2026.14

Comparative Analysis of Large Language Models in the Interpretation of Gynecologic Pathology Reports

Aslı Karakaşlı[1]

[1] Erol Olçok Training and Research Hospital, Department of Obstetrics and Gynecology, Çorum, Türkiye

OPEN ACCESS

CORRESPONDENCE
Aslı Karakaşlı
akarakasli.289@gmail.com

RECEIVED
11.12.2025
ACCEPTED
22.12.2025
PUBLISHED
05.01.2026

CITATION
Karakaşlı A. Comparative Analysis of Large Language Models in the Interpretation of Gynecologic Pathology Reports. *Eur J Innov Med Res.* 2026;1(1):31-33. doi:10.65495/eurjimr.2026.14

FINANCIAL SUPPORT
No external funding was received for this study.

CONFLICT OF INTEREST
The authors declare that they have no conflicts of interest related to this study.

ETHICAL APPROVAL
Not applicable.

INFORMED CONSENT
Not applicable.

ACKNOWLEDGEMENTS
None

PEER REVIEW
Reviewed by at least two peer-reviewers.

Dear Editor,

The increasing availability of digital health information has enabled patients to access their pathology reports prior to clinical consultation, a phenomenon that has markedly amplified what is described in the literature as "waiting-time anxiety".[1,2] In contemporary practice, patients increasingly rely on Large Language Models (LLMs) rather than conventional search engines to decipher complex medical terminology.[2–5] However, it remains uncertain how accurately and empathetically these models convey nuanced pathological concepts—particularly within "grey-zone" diagnoses such as Endometrial Intraepithelial Neoplasia (EIN), which cannot be classified as strictly benign or malignant—as well as how they communicate potential malignancy.[1–3,6] In this letter, we present a quantitative evaluation of the performance of three current LLMs across a series of gynecologic pathology scenarios.

In our study, we generated four synthetic pathology reports designed to represent a clinically relevant spectrum: benign (cellular leiomyoma), premalignant (endometrial intraepithelial neoplasia, EIN), indeterminate (atypical squamous cells of undetermined significance, ASC-US), and malignant (endometrioid adenocarcinoma) (Table 1). Three contemporary LLMs—Claude Sonnet 4.5, ChatGPT 5, and Gemini 3—were instructed to explain these reports to a persona defined as "a worried 45-year-old patient with no medical background." The resulting outputs were assessed using the Ateşman Readability Index, NRC Emotion Analysis, and a jargon-density metric.[7,8] The language of the study was Turkish. For readability assessment, the Ateşman Readability Index, which

Table 1. Synthetic Gynecologic Pathology Reports Used as Model Inputs

Case No.	Clinical Category	Pathology Report Text (Model Input)[1]
Case 1	Benign (Cellular Leiomyoma)	Gross Description: Nodular tissue fragment measuring 8 × 6 × 5 cm, with a beige–white cut surface containing focal cystic areas. Microscopic Description: Sections show intersecting fascicles of spindle-shaped smooth muscle cell bundles. Focal areas demonstrate edema, hyaline degeneration, and cystic change. Although mild increases in cellularity are noted in some regions, cytologic atypia is not prominent. No necrosis is identified. Mitotic activity is fewer than 1 per 10 HPF. Diagnosis: Cellular Leiomyoma, Uterus (Myomectomy Specimen).
Case 2	Premalignant (EIN)	Specimen: Endometrial Curettage. Microscopic Description: Examination of the entire specimen reveals increased glandular density with a gland-to-stroma ratio exceeding 1:1. Glands exhibit branching and crowding. Cytologic atypia is present, characterized by nuclear rounding, chromatin coarsening, and nucleolar prominence. The atypical glands are clearly distinguishable from the background endometrium. No evidence of invasion (myometrial involvement) is identified. Diagnosis: Findings Consistent with Endometrial Intraepithelial Neoplasia (EIN).
Case 3	Indeterminate (ASC-US)	Specimen: Cervical Smear. Microscopic Description: The background contains polymorphonuclear leukocytes and Döderlein bacilli. Superficial and intermediate squamous epithelial cells are present. Some squamous cells show nuclear enlargement (mildly increased nuclear-to-cytoplasmic ratio) and irregular nuclear contours; however, these findings are insufficient in quantity and quality to support a diagnosis of intraepithelial lesion (LSIL/HSIL). No cells suspicious for malignancy are identified. Diagnosis: Atypical Squamous Cells of Undetermined Significance (ASC-US).
Case 4	Malignant (Adenocarcinoma)	Specimen: Probe Curettage. Microscopic Description: Sections lack normal endometrial stroma. Instead, the tissue is replaced by back-to-back, cribriform, and complexly branching atypical glandular structures occupying the entire field. The neoplastic cells show marked nuclear pleomorphism, loss of polarity, and increased mitotic activity. A desmoplastic stromal reaction is present. Diagnosis: Endometrioid-Type Adenocarcinoma, FIGO Grade 1.

HPF: High-power field, EIN: Endometrial intraepithelial neoplasia, LSIL: Low-grade squamous intraepithelial lesion, HSIL: High-grade squamous intraepithelial lesion, ASC-US: Atypical squamous cells of undetermined significance, FIGO: International Federation of Gynecology and Obstetrics,
[1] *The content has been translated and adapted to comply with the journal's formatting and terminology guidelines.*

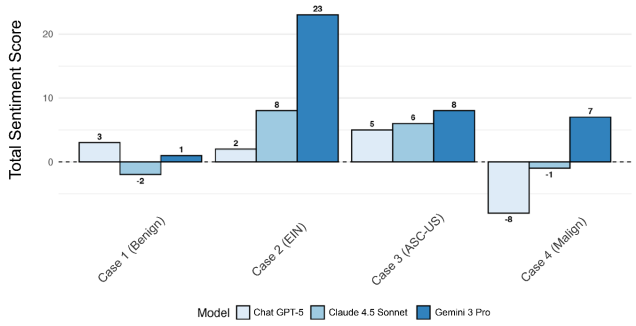

Figure 1. Sentiment Load of Model Responses (NRC Sentiment Score).

Higher scores indicate a greater density of positive or reassuring language, whereas lower scores reflect negatively valenced or alarming wording. The distribution across the four clinical scenarios (Benign, EIN, ASC-US, Malignant) illustrates substantial variation in emotional tone between models.

is specifically designed for and adapted to Turkish morphology, was employed. For sentiment analysis, the validated Turkish translation of the NRC Word-Emotion Association Lexicon (Saif Mohammad's NRC Word-Emotion Association Lexicon), accessible via the 'syuzhet' R package, was utilized.[9]

Our analyses revealed that none of the models adopted a standardized approach to patient education; instead, each demonstrated a distinct communicative profile (Table 2). Gemini 3 generated the longest and most detailed explanations (mean: 510 words) and incorporated the highest number of empathy markers (n=14), making it the model that conveyed the strongest empathetic intent (Figure 1).. However, its responses were heavily laden with technical terminology, resulting in markedly poor readability (mean Ateşman score: −99.9).

Conversely, Claude Sonnet 4.5 delivered the most balanced performance, offering concise yet adequately informative explanations (mean: 248 words). It achieved the highest readability scores in benign scenarios and, notably, eliminated potentially confusing terminology—such as "squamous" or "atypia"—in the ASC-US case, producing a fully jargon-free explanation (0.00% jargon density). In the EIN scenario, its use of the metaphor "This is not a red light but a yellow one" to describe diagnostic uncertainty was identified as an exemplary strategy for reducing patient anxiety.

Although ChatGPT 5 demonstrated a high degree of technical accuracy, it consistently underperformed in the domain of "emotional intelligence." In three of the four scenarios, the model produced responses entirely devoid of empathy markers. More importantly, in the malignant scenario, its use of starkly negative language failed to incorporate the essential buffering and softening strategies emphasized in established "breaking bad news" protocols.

Taken together, our findings suggest that these three models assume distinct functional roles from the patient's perspective: Gemini 3 resembles an "Academic Instructor" that appeals to detail-oriented users; ChatGPT 5 functions more as a detached "Technical Glossary"; and Claude Sonnet 4.5 operates as an "Empathic Clinician" with a focus on anxiety mitigation. Ultimately, our results illustrate the diverse communicative profiles patients may encounter when independently consulting these tools. Clinicians' awareness of these varying "AI communication styles" is critical—not only for correcting unrealistic patient expectations but also for managing secondary anxiety that may arise from digital information overload.

References

1. Steimetz E, Minkowitz J, Gabutan EC, et al. Use of Artificial Intelligence Chatbots in Interpretation of Pathology Reports. JAMA Netw Open. May 22 2024;7(5):10. e2412767. doi:10.1001/jamanetworkopen.2024.12767

2. Beale SK, Cohen N, Secheli B, McIntire D, Kho KA. Comparing physician and artificial intelligence chatbot responses to posthysterectomy questions posted to a public social media forum. AJOG Glob Rep. Aug 2025;5(3):11. 100553. doi:10.1016/j.xagr.2025.100553

3. Anastasio MK, Peters P, Foote J, et al. 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. Oct 2024;55:4. 101477. doi:10.1016/j.gore.2024.101477

4. Kowalski JT, Brechtel L. Review of chatbots in urogynecology. Curr Opin Obstet Gynecol. Dec 2025;37(6):421-425. doi:10.1097/gco.0000000000001067

5. Recker F, Neubauer R, Wittek A, Scholten N. Large language models and women's health: a digital companion for informed decision-making. Editorial Material. Arch Gynecol Obstet. Sep 2025;312(3):663-670. doi:10.1007/s00404-025-08065-9

Table 2. Quantitative Comparison of LLM Outputs Across Four Gynecologic Pathology Scenarios

Model	Clinical Scenario	Word Count	Readability[1]	Sentiment Score[2]	Jargon Density[3] (%)	Lexical Diversity[4]
Claude Sonnet 4.5	Case 1 (Benign)	208	−51.41	−2	2.54	0.75
	Case 2 (Premalign)	240	−81.08	8	0.90	0.81
	Case 3 (ASC-US)	299	−77.90	6	0.00	0.70
	Case 4 (Malignant)	245	−79.56	−1	0.43	0.77
ChatGPT 5	Case 1 (Benign)	280	−63.88	3	1.52	0.71
	Case 2 (Premalign)	305	−75.38	2	0.35	0.76
	Case 3 (ASC-US)	292	−76.02	5	0.00	0.76
	Case 4 (Malignant)	380	−92.01	−8	2.16	0.64
Gemini 3	Case 1 (Benign)	458	−98.51	1	2.43	0.64
	Case 2 (Premalign)	496	−98.08	23	1.01	0.70
	Case 3 (ASC-US)	514	−100.00	8	1.78	0.64
	Case 4 (Malignant)	573	−89.86	7	1.60	0.63

Data were analyzed using R (v4.3.1). [1] Ateşman Readability Formula: The Turkish adaptation of the Flesch Reading Ease method; higher scores (i.e., values approaching zero) denote greater readability. Negative values are expected in medically technical content due to the high density of specialized terminology. [2]Sentiment Score: Calculated using the NRC sentiment lexicon. Negative scores indicate alarmist or negatively valenced language, positive scores reflect reassuring or supportive language, and a score of 0 denotes a neutral tone. [3]Jargon Density: The proportion of predefined medical terms (e.g., "neoplasia," "atypia") relative to the total word count. Lower proportions indicate greater patient-centered simplification. [4]Lexical Diversity (Type–Token Ratio, TTR): The ratio of unique word types to the total number of words (range: 0–1). Higher values reflect reduced word repetition and greater linguistic richness.

6. Cohen ND, Ho M, McIntire D, Smith K, Kho KA. A comparative analysis of generative artificial intelligence responses from leading chatbots to questions about endometriosis. AJOG Glob Rep. Feb 2025;5(1):7. 100405. doi:10.1016/j.xagr.2024.100405

7. Ateşman E. Türkçede Okunabilirliğin Ölçülmesi. Measuring readability in Turkish. Dil Dergisi. 1997;(58):71-74.

8. Mohammad SM, Turney PD. NRC emotion lexicon. National Research Council of Canada. Record identifier / Identificateur de l'enregistrement : 0b6a5b58-a656-49d3-ab3e-252050a7a88c, Collection / Collection : NRC Publications Archive / Archives des publications du CNRC. Updated 2013/11/15. https://nrc-publications.canada.ca/eng/view/object/?id=0b6a5b58-a656-49d3-ab3e-252050a7a88c

9. Jockers ML. Syuzhet: Extract Sentiment and Plot Arcs from Text. Accessed 21 November, 2025. https://github.com/mjockers/syuzhet