EFC XI EUROPEAN CONGRESS
European Federation for Colposcopy and Cervical Cancer Prevention
Bucharest, Romania | 14-17 October 2026

# Can Large Language Models (LLMs) make colposcopy easy to read? An exploratory case study on simplified colposcopic reports

Natalini Leonardo[1], Bertuzzi Marta[1], Romagnoli Elisa[1], Ulisse Chiara[1], Delli Carpini Giovanni[1], Ciavattini Andrea[1]
[1]Obstetrics and Gynecologic Section, Department of Odontostomatologic and Specialized Clinical Sciences, Università Politecnica delle Marche, Ancona, Italy

Introduction: Colposcopic reports are technical and hard for patients to understand. Large Language Models (LLMs) could simplify them, but their reliability for this task is unknown. We assessed the content quality and readability of LLM-simplified colposcopic reports, and whether performance varied with case severity.

Methods: Ten fictitious colposcopic reports of increasing case severity (from normal to carcinoma) were simplified by six LLMs (ChatGPT, Gemini, Qwen, Claude Sonnet, Llama, Mistral). Three blinded colposcopists rated overall quality (GQS), understandability (PEMAT-U), accuracy (mDISCERN), completeness (CP), and, where present, error severity (ES). Readability was computed (Flesch Reading Ease [FRE], SMOG [Simple Measure of Gobbledygook], Dale-Chall readability formula [DC], reading time [RT]). Between-model differences and severity trends were tested.

Results: Content quality was high, with median GQS 4.00 [4.00-5.00], PEMAT-U 87.00 [80.00-93.00], mDISCERN 4.00 [3.00-4.00], and CP 5.00 [4.00-5.00], and comparable across models except for Llama, which scored lower on every measure. Readability level was moderate to low (mean FRES 55.96 ± 8.14, SMOG 11.43 ± 1.28, DC 10.88 ± 0.63, RT 97.89 ± 28.20), differed between models ($p<0.001$) and worsened as severity increased. Errors were flagged in 14/60 responses (23%), 8 of them (57%) from Llama, and were all minor or moderate (ES median 4, IQR 4-5). No critical or life-threatening error occurred.

Conclusions: LLM-simplified colposcopic reports showed high content quality and few, minor errors, but remained moderately difficult to read, more so in severe cases. LLMs may serve as an adjunct to expert counselling.

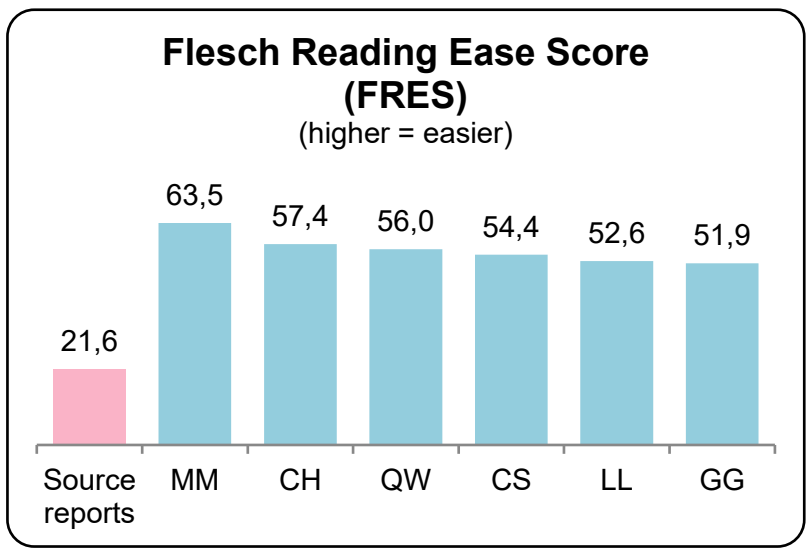


Figure 1. Mean FRES of the ten source colposcopic reports and of the responses generated by each LLM. Higher scores indicate easier text.
CH, GPT-5.5 Instant; GG, Gemini 3.6 Flash; LL, Llama 4 Scout; MM, Mistral Medium 3.5; QW, Qwen3.7-Max; CS, Claude Sonnet 5.

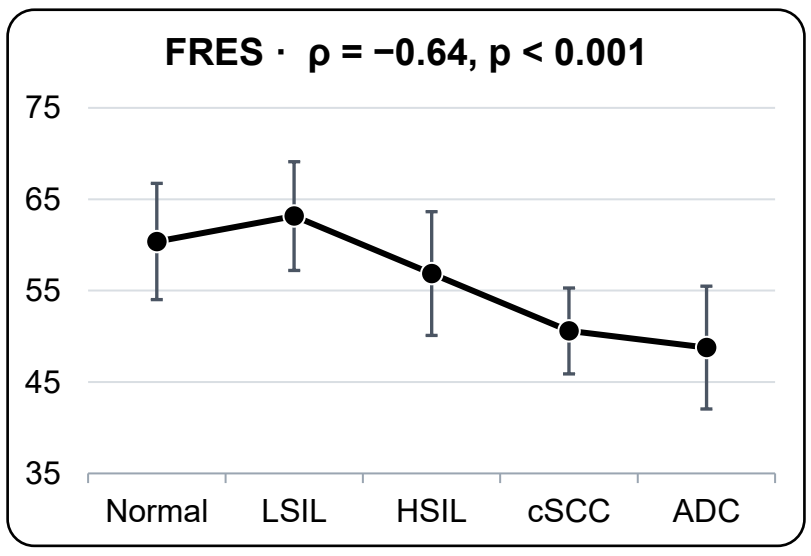


Figure 2. FRES of the LLM responses by severity of the source report. Readability decreased progressively as severity increased.
LSIL/HSIL, low-/high-grade squamous intraepithelial lesion; cSCC, cervical squamous cell carcinoma; ADC, adenocarcinoma.