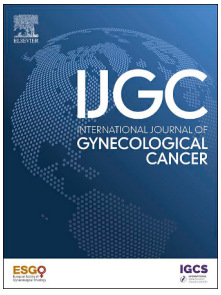
ORIGINAL RESEARCH

Natural language processing as consultation service platform or clinical decision support system in gynecologic oncology: a systematic review



Andrea Rosati[a,b,1,*], Massimo Criscione[b,c,1], Livia Lilli[b,d], Marco Petrillo[c], Giampiero Capobianco[c], Stefano Patarnello[d], Anna Fagotti[a,b]

Received 8 October 2025, Accepted 1 February 2026; Available online 7 February 2026

ABSTRACT

Objective: Natural language processing is emerging as a key application of artificial intelligence in oncology. This systematic review aims to evaluate the performance and methodological frameworks of natural language processing systems in gynecologic oncology.

Methods: We conducted a systematic review following the Preferred Reporting Items for Systematic Reviews and Meta-Analysis 2020 guidelines. MEDLINE, EMBASE, and Web of Science were searched for studies published between January 2015 and February 2025. Outcomes were synthesized across 3 research questions: the accuracy of natural language processing systems used as consultation service platforms; the accuracy of natural language processing systems used as clinical decision support systems; and the benchmarking methodologies applied, including their associated methodological outcomes. Consultation service platforms deliver general medical information, whereas clinical decision support systems provide recommendations that are integrated into the patient's clinical workflow.

Results: This review analyzed 12 retrospective studies. Consultation service platforms were less accurate than clinicians (60% vs 86.7%) and rated lower in response quality (2.96/5 vs 4.2/5) but outperformed guideline-based answers (1.54/2 vs 1.38/2). In cervical cancer, ChatGPT surpassed experts (7.0 vs 6.1). ChatGPT-4 showed a concordance of 70% with the National Comprehensive Cancer Network and 60% with the European Society of Gynaecological Oncology guidelines in clinical decision support tasks, with an overall recommendation accuracy of 75%. IBM Watson achieved a 72.8% concordance with guidelines. Prompting was applied from 100% to 37.5% across studies. Qualitative benchmarking varied across studies: 83.3% used clinical guidelines and 37.5% of consultation service platforms studies used expert answers. Four- or 5-point scales and binary scoring were used to assess consultation service platforms and clinical decision support systems, respectively.

Conclusions: Clinicians remain superior in complex reasoning, but natural language processing systems demonstrate robust performance in guideline-driven tasks, with advantages in speed, readability, and reproducibility. However, performance declined in nuanced scenarios and among under-represented patient sub-groups. Large language models currently play a supportive rather than substitutive role in gynecologic oncology.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Natural language processing systems are increasingly applied in medicine, including oncology. However, their role in gynecologic oncology remains poorly defined, with fragmented evidence on their accuracy and methodological robustness when used as consultation platforms or clinical decision support systems.

WHAT THIS STUDY ADDS

This systematic review synthesizes 12 studies evaluating natural language processing in gynecologic oncology, showing that although clinicians remain superior in complex reasoning, natural language processing systems achieve strong performance in structured, guideline-driven contexts. They offer advantages in speed, readability, and reproducibility, although performance varies by clinical domain and task complexity.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE, OR POLICY

The findings support the cautious integration of natural language processing into gynecologic oncology as an adjunct rather than a replacement for clinicians. Future research should prioritize standardized evaluation protocols, validation in complex scenarios, and improved transparency to ensure safe and equitable deployment in clinical practice.

* ***Correspondence to*** *Dr Andrea Rosati, Department of Woman's and Child Health and Public Health Sciences, Gynecologic Oncology Unit, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy; dott.andrearosati89@gmail.com (A. Rosati), massimo.crisc@gmail.com (M. Criscione), livia.lilli@policlinicogemelli.it (L. Lilli), marco.petrillo@gmail.com (M. Petrillo), capobia@uniss.it (G. Capobianco), stefano.patarnello@gemelligenerator.it (S. Patarnello), anna.fagotti@policlinicogemelli.it (A. Fagotti)*

[1] AR and MC contributed equally to this work.

https://doi.org/10.1016/j.ijgc.2026.104558

Keywords:
Large Language Models; Natural Language Processing; Clinical Decision Support System; Consultation Service Platform; Tumor Board

ORIGINAL RESEARCH

INTRODUCTION

Natural language processing is a sub-field of artificial intelligence that leverages deep learning and rule-based methodologies to process and generate human language. Large language models represent the forefront of natural language processing technologies in clinical practice that can be categorized by pre-training domain (eg, general, medical, reasoning, math, multi-lingual), model architecture (encoder, decoder, or encoder-decoder), and parameter status (frozen vs unfrozen). Among general-purpose large language models, OpenAI's GPT series (eg, ChatGPT), using a decoder-only architecture, are more efficient in text generation,[1] whereas Google's encoder-only BioBERT performs better in text understanding and decoding.[2] Other models, such as Meta's LLaMA and Google Health's MedPaLM, reflect a growing trend toward clinically specialized large language models, each with tailored architecture and pre-training domains.[3,4] From an accessibility perspective, open-source models such as Mistral and DeepSeek are gaining traction due to their deployability on local infrastructure and unfrozen status.[5-7] A sub-group of large language models is chatbot-based. More details on large language models commonly used in health care are available in Table S1.

In parallel, non—large language model natural language processing systems, such as IBM Watson for Oncology, which integrate rule-based artificial intelligence with traditional algorithms, have been used as clinical decision support tools across various cancers.[8] From a functional standpoint, natural language processing applications in health care fall into 3 main domains (Table S2).

In clinical text processing and information extraction, natural language processing systems extract structured data from unstructured texts such as physician notes and discharge summaries, improving documentation and reducing administrative workload, which accounts for up to 25% of clinicians' time.[9-11] Consultation service platforms are natural language processing systems that provide medical information to patients and clinicians, supporting education and knowledge dissemination without relying on patient-specific data.[12,13] Clinical decision support systems are natural language processing systems that encode guidelines or learn clinical reasoning from data sets.[14-17] They generate tailored recommendations, validated against expert benchmarks, before integration into multi-disciplinary workflows.[18,19]

The primary objective of this systematic review is to evaluate the performance and methodological quality of natural language processing systems applied in gynecologic oncology as consultation service platforms or clinical decision support systems.

METHODS

The protocol of this systematic review is registered on the International Prospective Register of Systematic Reviews, registration number CRD42024589109. Retrospective and prospective studies were included. For natural language processing as a clinical decision support system, the evaluated population comprised patients with gynecologic malignancies without restrictions on age, histology, International Federation of Gynecology and Obstetrics stage, body mass index, Eastern Cooperative Oncology Group status, or sample size.

The following types of studies were excluded: reviews, systematic reviews, discussion papers, non-research editorials, qualitative studies, conference abstracts, animal studies, and gray literature.

Using the PubMed interface for MEDLINE, we developed a search strategy (Table S3), adapted for EMBASE and Web of Science, to retrieve relevant studies. We included published and unpublished data, restricted to English-language articles. On February 26, 2025, we systematically searched MEDLINE, EMBASE, and Web of Science, targeting titles and abstracts. ClinicalTrials.gov was also searched using the same strategy. Articles published between January 2015 and February 2025 were included.

To identify eligible studies, 2 reviewers (AR and MC) independently screened titles and abstracts, followed by full-text reviews of potentially relevant articles. Discrepancies were resolved through discussion and consensus; a third reviewer (AF) was consulted when needed. Missing data were addressed by contacting study authors. The screening tool Rayyan was used to remove duplicates and manage title/abstract screening. The selection process is illustrated in Figure 1 in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) 2020 statement (the PRISMA 2020 flow diagram was generated using https://estech.shinyapps.io/prisma_flowdiagram/).[20,21] The PRISMA checklist is available in Table S9.

Two reviewers (AR and MC) independently assessed the quality of the included studies and the overall systematic review. Risk of bias was evaluated using a modified version of the Joanna Briggs Institute Critical Appraisal Checklist for Qualitative Research (2015),[22] with results presented in Table S4 and graphical plots (Figs. S5-S6). The synthesis and summary of qualitative findings were conducted using the Grading of Recommendations Assessment, Development, and Evaluation - Confidence in the Evidence from Reviews of Qualitative research (GRADE-CERQual) approach.[23] The GRADE-CERQual checklist is available in Supplementary Material 4.

For data extraction, 2 authors independently reviewed all eligible articles. Index data are provided in Table S7. Descriptive tables were used to summarize the relevant data from the studies included. The study data were summarized into tables using a predefined checklist.

The results of the data extraction process were primarily synthesized using a narrative approach, wherein the key findings were systematically presented in the study. An accompanying glossary (Table S8) defines essential terms relevant in this systematic review. Artificial intelligence engineers with expertise in large language models (SP and LL) and statistical programming were involved, focusing on the evaluation metrics, data set scale and characteristics, and the types of language models assessed across studies.

The review was conducted according to the following 3 established research questions (Fig. 2). Research question 1: How accurate are natural language processing systems in answering questions related to gynecological oncology when used as a consultation service platform? Research question 2: How accurate

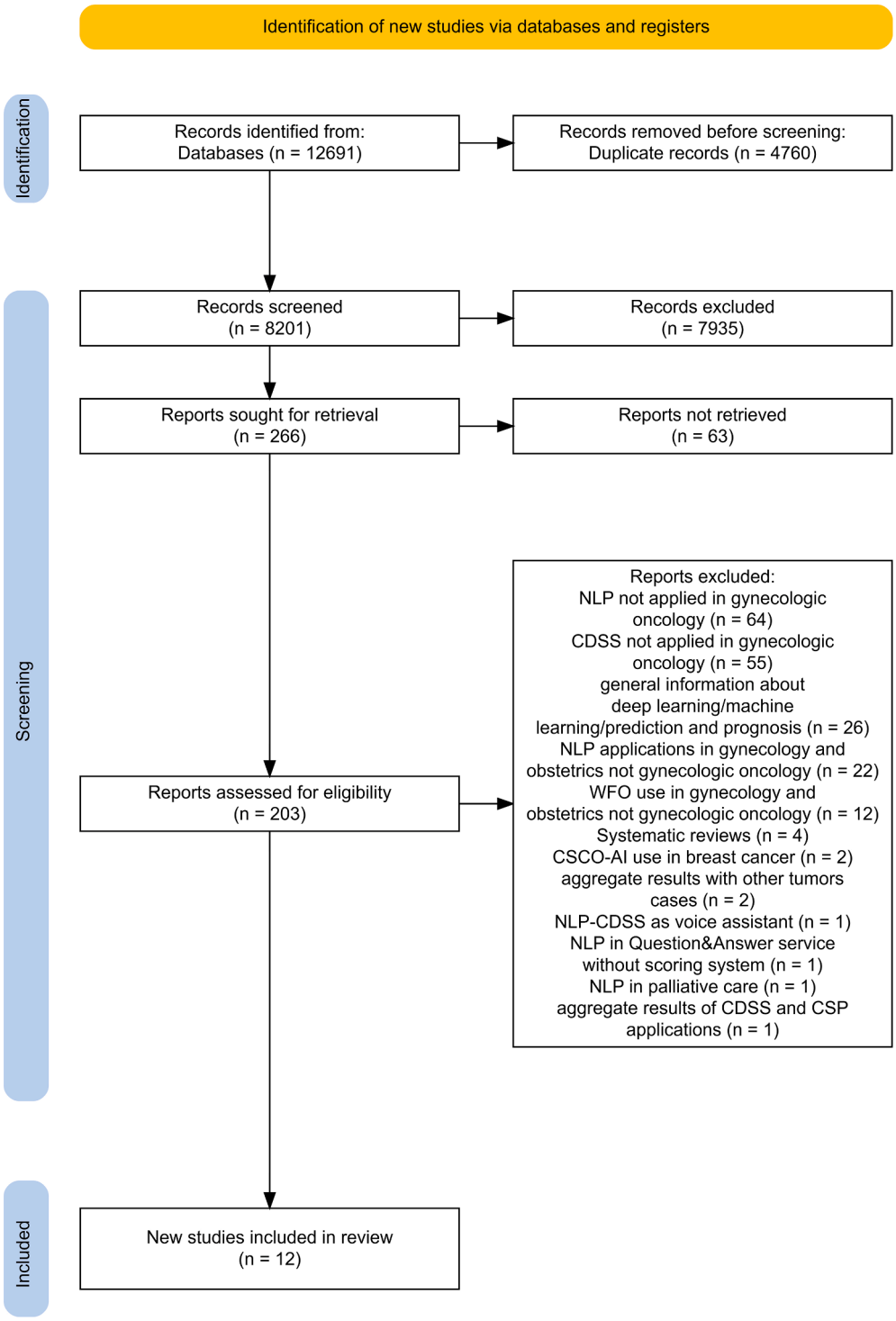


Figure 1 Adapted Preferred Reporting Items for Systematic Reviews and Meta-Analysis Guidelines 2020 flow diagram (based on the study by Page and colleagues).[21]

are natural language processing systems in suggesting appropriate treatments for patients with gynecological tumors when used as a clinical decision support system? Research question 3: Which methodological frameworks are used to control interaction bias, and how is model performance benchmarked qualitatively and quantitatively for consultation service platforms and clinical decision support systems?

This study is a systematic review of publicly available literature and did not involve human participants; therefore, institutional review board approval and informed consent were not required.

In accordance with the journal's guidelines, we will provide our data for independent analysis by a selected team by the Editorial Team for the purposes of additional data analysis or for the reproducibility of this study in other centers, if such is requested.

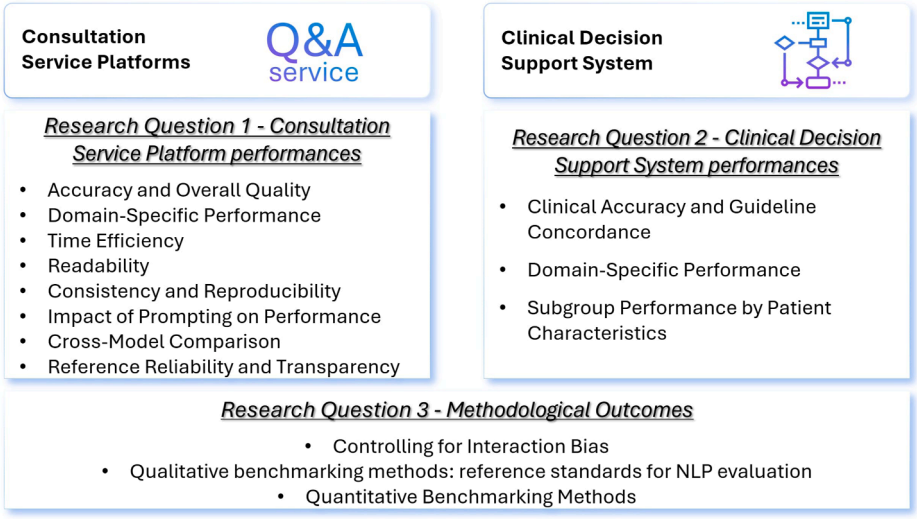


Figure 2 Research questions.

RESULTS

The literature search identified 12,691 records: EMBASE (n = 2832), MEDLINE (n = 1965), and Web of Science (n = 8164). After removing 4760 duplicates, 8201 articles were screened by title and abstract. Of these, 63 reports were not retrieved, and 203 full-text articles were excluded. Ultimately, 12 studies met the inclusion criteria and were included in the analysis. No relevant clinical trials were found in the ClinicalTrials.gov (National Institutes of Health) registry.

The main characteristics of the included studies are presented in Tables S10 and S11.

Natural language processing applications in gynecologic oncology have been evaluated in terms of performance and methodological outcomes across 12 retrospective studies, including 8 on consultation service platforms (Table S10) and 4 on clinical decision support systems (Table S11).

In 91.7%, proprietary general-purpose large language models were used, most commonly ChatGPT (83.3%), with 1 study assessing the rule-based Watson for Oncology (8.3%).

Cervical cancer was the most frequently investigated (66.7%), followed by ovarian (50.0%), endometrial (33.3%), and genetic counseling (8.3%).

Consultation service platform studies evaluated between 8 and 127 clinical questions, whereas clinical decision support system studies assessed a total of 292 patient cases, with sample sizes per study ranging from 10 to 246.

No quantitative synthesis was performed; therefore, formal statistical assessment of reporting bias was not applicable.

Regarding research question 1, which focused on the performance outcomes of natural language processing systems as a consultation service platforms, Anastasio and colleagues[24] found that expert clinicians showed higher factual accuracy (86.7% vs 60%) and superior perceived response quality (4.2 vs 2.96 on a 5-point Likert scale). Nonetheless, in the study by Finch and colleagues[25], GPT-4 demonstrated a modest advantage over traditional guideline-derived answers in ovarian cancer management, achieving higher completeness and accuracy; GPT-4 responses scored 1.54 (0 to 2 accuracy scale), compared with 1.38 for National Comprehensive Cancer Network (NCCN)—based output. Concerning domain-specific performances, in narrowly defined and structured clinical scenarios such as cervical cancer frequently asked questions, large language models exceeded expert benchmarks: Ye and colleagues[26] reported a significantly higher median score for ChatGPT (7.0) compared with human experts (6.1) (7-point accuracy scale) (Table S10).

Hermann and colleagues[27] reported high accuracy of large language models for prevention (91.7%), survivorship, and quality of life (93.8%), with a drop to 33.3% for diagnostic tasks and 71.4% for treatment-related responses. Similarly, in the study by Patel and colleagues,[28] GPT-4 achieved 100% accuracy in general genetics, falling to 70.6% for hereditary breast and ovarian cancer and 66.6% for Lynch syndrome (Table S10). In cervical cancer, Yurtcu and colleagues[29] found that 91.9% of responses to frequently asked questions were rated “excellent” (global quality score 5), with a drop to 62.3% for specific queries based on the European Society of Gynaecological Oncology (ESGO)/European Society for Radiotherapy and Oncology/European Society of Pathology guidelines (Table S10).

Large language models demonstrated advantages in response time and textual output compared with clinicians. Specifically, ChatGPT provided faster responses (mean 15 seconds per answer vs 24 seconds for clinicians) and generated longer outputs (mean 307.8 vs 252.6 words), as reported by Anastasio and colleagues.[24] In terms of readability, ChatGPT achieved higher Flesch reading ease scores (40.62 vs 32.52), indicating easier-to-read text, whereas the Flesch—Kincaid grade level scores were comparable between ChatGPT and clinicians (12.26 vs 12.66), suggesting similar levels of textual comprehension, as reported by Ye and colleagues[26] (Table S10). The reproducibility rates of ChatGPT were 93.2% for frequently asked questions and 88.7% for cervical cancer—specific guideline-based queries across repeated submissions in the study by Yurtcu and colleagues.[29]

Similarly, Ye and colleagues[26] confirmed stable and consistent outputs across multiple sessions conducted at different times and locations.

In the study by Kuerbanjiang and colleagues,[30] prompting improved accuracy from 2.52 to 2.67 with ChatGPT-4.0 Turbo, whereas Gemini Pro increased from 2.00 to 2.25 (66% to 77%). In

contrast, Claude 2 showed virtually no change (2.39 unprompted vs 2.35 prompted; both 85%), Phi2 remained flat (1.44 vs 1.50; 49%), and HuatuoGPT declined (1.76 to 1.56; 57% to 53%). Finch and colleagues[25] reported a 2% concordance improvement with prompting, whereas Piazza and colleagues[31] found no statistically significant difference between GPT-4 prompted (2.38) and unprompted (2.49) formats (Table S10).

Ye and colleagues[26] found that only 50% of ChatGPT's citations were accessible through verifiable sources. Yurtcu and colleagues,[29] Piazza and colleagues,[31] and Hermann and colleagues,[27] reported use of vague or non-specific references attributed to clinical guidelines without precise bibliographic references (Table S10).

Exploring research question 2, namely the performance outcomes of natural language processing systems as clinical decision support system; Levin and colleagues[32] reported concordance rates of 70% between ChatGPT-4 and NCCN guidelines and 60% with ESGO guidelines, with an overall recommendation accuracy of 75%. In the study by Ebner and colleagues,[33] when benchmarked against NCCN guidelines, ChatGPT-3.5 reached a mean concordance score of 1.7 (0 to 6 accuracy scale). Zou and colleagues[34] showed a 72.8% concordance between IBM Watson for Oncology and NCCN guidelines in cervical cancer cases. Regarding domain-specific performance, Levin and colleagues[32] reported that ChatGPT-4 achieved a concordance rate of 90% in ovarian cancer—related scenarios, whereas performance was substantially lower in endometrial cancer, with a concordance of 20%. Krückel and colleagues[35] reported the opposite trend, with endometrial cancer receiving the highest average guideline concordance score (1.5) and lower values for ovarian and cervical cancers (0.75). The results from Zou and colleagues[34] showed a significant drop in concordance among older patients (age $\geq$65 years: 38.5%; odds ratio [OR] 0.08, 95% confidence interval [CI] 0.03 to 0.28, $p = .03$), those with poorer functional status (Eastern Cooperative Oncology Group $\geq$3: 23.1%; OR 0.29, 95% CI 0.083 to 1.058, $p = .05$), and early (stage I: 41.4%) and stage IV disease (stage IV: 7.6%) (Table S11).

We then analyzed research question 3, which explores the methodological outcomes of natural language processing systems when applied as consultation service platforms and as clinical decision support systems.

Prompting was reported with substantial variability across studies (from 100% to 37.5%). Prompted vs unprompted answers were directly compared in 3 of 8 studies on consultation service platforms (Table S10), whereas all studies on clinical decision support systems implemented prompting (Table S11). To control interaction bias, single-turn "fresh session" protocols were adopted in 6 consultation service platform studies (Table S10) and in all clinical decision support system studies (Table S11). Repetition for consistency testing was reported in 2 consultation service studies, using repeated submissions across different sessions, time points, or locations to assess intra-model stability (Table S10). Where formally assessed, the impact of prompting on performance was model dependent (Table S10).

As qualitative benchmarking methods, 3 types of references were used to evaluate natural language processing performance. Ten studies used established clinical guidelines, such as NCCN, ESGO/European Society for Radiotherapy and Oncology/European Society of Pathology, Associazione Italiana di Oncologia Medica, and German national guidelines, as primary reference standards for performance evaluation. This included 75% of consultation service platform studies (Table S10) and all clinical decision support system studies (Table S11). Expert-derived answers were used to assess responses in 37.5% of consultation service platform studies responses (Table S10). Finally, 75% of clinical decision support system studies compared natural language processing performance with real-world clinical decisions made by multi-disciplinary tumor boards (Table S11).

Quantitative benchmarking methods varied across studies: among consultation service platform studies, structured rating systems were used. One study used binary accuracy; 1 study used a 3-point scale, 3 studies used a 4-point scale (effectiveness or completeness); 3 studies used a 5-point Likert or global quality score scale; and 1 study used a 7-point scale assessing accuracy, relevance, and consistency (Table S10). In clinical decision support system studies, concordance metrics included binary concordance in 2 studies and a 4-point and a 7-point scale in 1 study each, respectively (Table S11).

DISCUSSION

Summary of Main Results

As a consultation service platform, natural language processing systems outperformed guideline-based responses and expert clinicians in structured tasks, supporting their role as adjunctive tools in standardized contexts. However, their performance declines in more complex or less codified scenarios.[25,26] Prompting represents an adaptation strategy to optimize model performance; however, its effectiveness is model dependent and often plateaus in high-performing architectures. Although large language models tend to generate longer outputs that could enhance patient comprehension, their use requires optimization in clinical contexts where concise communication is beneficial for decision-making purposes. Finally, source reliability remains a major concern because hallucinated references, reported in up to 50% of cases,[26] undermine credibility and explainability. As a clinical decision support system, natural language processing holds a dual potential: to promote standardization by aligning clinical decisions with current guidelines and enable personalization within a precision medicine framework.

Large language models have shown growing concordance with oncology standards, achieving rates from 60% to 90%. Yet, significant discordance persists: 25% for ChatGPT-4 and 27.2% for Watson for Oncology, particularly in nuanced or non-standardized clinical contexts.[32,34]

Performance also varied both by tumor type, with 90% concordance in ovarian cancer vs 20% in endometrial cancer,[32] whereas another author reported the highest mean concordance for endometrial cancer (1.5), with lower values for ovarian and cervical cancers (0.75).[35] This divergence likely reflects methodological rather than biological differences, including different guideline anchors (NCCN/ESGO vs German national guideline), different model capability and versioning (ChatGPT-4 vs ChatGPT-3.5), case-mix variation, and scenario granularity (eg, early-stage endometrial scenarios may require higher risk stratification, adjuvant tailoring, and molecular integration). Finally, our findings reveal considerable heterogeneity in strategies used to control interaction bias and benchmarking approaches across studies.

Results in the Context of Published Literature

The results of this systematic review are consistent with findings reported in broader oncologic and medical domains, where natural language processing and large language models have shown promising but variable performance depending on task structure and clinical complexity. In line with previous studies, large language models demonstrated higher accuracy in guideline-driven contexts but poorer performance in individualized or multi-modal decision-making scenarios.[19,36]

In previous evaluations across other specialties, concordance between large language models and clinical guidelines ranged from 60% to 85%. Similar to our findings, Watson for Oncology and GPT-based systems often performed well for standardized treatment algorithms but struggled with rare histotypes or patient co-morbidities.[37]

Regarding patient-facing applications, previous research in breast, gastric, and hepatobiliary cancers has confirmed that natural language processing-based consultation tools improve the accessibility and readability of medical information. However, they remain limited by factual inaccuracies and incomplete referencing.[13,38] These trends were mirrored in our analysis, where citation reliability was frequently sub-optimal, thereby undermining the explainability of the models.

Our findings identify prompt engineering and bias control as critical determinants of model reliability. Previous studies in radiology and internal medicine have demonstrated that prompt optimization can improve accuracy by approximately 10% to 15%, in line with the modest yet reproducible performance gains observed in the present analysis.[39]

Strengths and Weaknesses

To the best of our knowledge, this is the first systematic review addressing the application of natural language processing in gynecologic oncology. The strengths of this analysis include the inter-disciplinary evaluation conducted by clinicians and engineers, adherence to the PRISMA checklist, formal risk of bias assessment, and appraisal of evidence quality using the GRADE-CERQual framework.[21,23] The current evidence base is limited by its retrospective design, small sample sizes, and methodological heterogeneity. The lack of standardized benchmarking tools further prevented a robust quantitative synthesis. A further limitation is the impossibility of critically comparing model architectures and training data sets because most systems are proprietary and closed-source; this constraint lies beyond the authors' control and limits the transparency of cross-model comparisons. These factors restrict the generalizability of our findings and emphasize the need for prospective, rigorously designed studies.

Implications for Practice and Future Research

From the patient's perspective, consultation service plat form—enabled conversations support education, empowerment, and adherence to cancer screening,[40-42] whereas natural language processing systems underpin telemedicine by facilitating remote monitoring, rapid assistance, psychological support, and care coordination.[43,44] From the clinician's perspective, natural language processing enhances standardization and personalization of care, providing real-time evidence-based guidance that promotes equity across settings and improves efficiency and cost-effectiveness in oncologic workflows. A key advantage is high time efficiency and reduced clinical workload.[9] Standardized reporting frameworks, such as a TRIPOD checklist for large language model—based studies, are needed to ensure transparency in model design, data provenance, human oversight, prompt engineering, and task-specific metrics, thereby supporting reproducibility and safe clinical deployment.[40]

The growing emphasis on model transparency and adaptability supports the development of open, unfrozen, fine-tunable conversational systems for secure in-house use. These systems should be co-developed and validated by multi-disciplinary tumor boards of clinicians and artificial intelligence experts to deliver flexible, context-aware decision support.[39]

Future research should prioritize enhancing guideline concordance, refining prompting strategies, integrating retrieval-augmented generation, and developing hybrid systems that merge artificial intelligence outputs with real-time expert oversight in a dynamic, continuously learning framework.

CONCLUSIONS

This systematic review highlights the potential and current limitations of natural language processing systems in gynecologic oncology. Although human clinicians remain superior in diagnostic reasoning and interpretive tasks, large language models demonstrated strong performance in structured, guideline-driven contexts, occasionally surpassing experts. However, performance remains highly dependent on clinical domain and task complexity, and the high prevalence of unverifiable or fabricated citations further undermines trust in clinical outputs. A major finding is the lack of standardized benchmarking, underscoring the need for locked guideline anchors (eg, NCCN/ESGO/national) and harmonized metrics capturing guideline concordance and clinical appropriateness. In conclusion, large language models currently play a supportive rather than substitutive role in gynecologic oncology.

Author Affiliations

[a]Fondazione Policlinico Universitario A. Gemelli, IRCCS, Woman, Child and Public Health Department, Roma, Italy
[b]Università Cattolica del Sacro Cuore, Rome, Italy
[c]University of Sassari, Department of Medicine, Surgery and Pharmacy, Sassari, Italy
[d]Fondazione Policlinico Universitario A. Gemelli, IRCCS, Rome, Italy

Funding/Support This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contributions AR contributed to the concept and design of the review. AR, LL, and MC analyzed and interpreted the included data. MC contributed to data collection. SP contributed to drafting the manuscript. AF, MP, and GC critically revised the manuscript for important intellectual content. All authors read and approved the final manuscript.

Declaration of Competing Interests None declared.

Data Availability All data supporting the findings of this study are included in the article and its supplementary information files; further details are available from the corresponding author upon reasonable request.

Appendix A. Supplementary data Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijgc.2026.104558.

REFERENCES

1. Achiam OJ. GPT-4 Technical Report. arXiv. Accessed February 26, 2025. https://arxiv.org/abs/2303.08774
2. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234—1240. https://doi.org/10.1093/bioinformatics/btz682.
3. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: toward building open-source language models for medicine. *J Am Med Inform Assoc*. 2024;31(9):1833—1843. https://doi.org/10.1093/jamia/ocae045.
4. Merlino DJ, Brufau SR, Saieed G, et al. Comparative assessment of otolaryngology knowledge among large language models. *Laryngoscope*. 2025;135(2):629—634. https://doi.org/10.1002/lary.31781.
5. Temsah A, Alhasan K, Altamimi I, et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. *Cureus*. 2025;17(2):e79221. https://doi.org/10.7759/cureus.79221.
6. Vaid A, Duong SQ, Lampert J, et al. Local large language models for privacy-preserving accelerated review of historic echocardiogram reports. *J Am Med Inform Assoc*. 2024;31(9):2097—2102. https://doi.org/10.1093/jamia/ocae085.
7. Zhong J, Sehgal K, Hickey K, et al. A local large language model pipeline automatically risk stratifies pancreatic cysts for population health management from serial radiology reports. *Gastroenterology*. 2024;166:S-687. https://doi.org/10.1016/s0016-5085(24)02038-9.
8. Yue L, Yang L. Clinical experience with IBM Watson for Oncology (WFO) for multiple types of cancer patients in China. *Ann Oncol*. 2017;28:x162. https://doi.org/10.1093/annonc/mdx676.024.
9. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med (Lond)*. 2023;3(1):141. https://doi.org/10.1038/s43856-023-00370-1.
10. Tavabi N, Pruneski J, Golchin S, et al. Building large-scale registries from unstructured clinical notes using a low-resource natural language processing pipeline. *Artif Intell Med*. 2024;151:102847. https://doi.org/10.1016/j.artmed.2024.102847.
11. Zeng K, Pan Z, Xu Y, Qu Y. An ensemble learning strategy for eligibility criteria text classification for clinical trial recruitment: algorithm development and validation. *JMIR Med Inform*. 2020;8(7):e17832. https://doi.org/10.2196/17832.
12. Das R, Maheswari K, Siddiqui S, et al. Improved precision oncology question-answering using agentic LLM. medRxiv. February 26, 2025. https://www.medrxiv.org/content/10.1101/2024.09.20.24314076v1
13. Kim AR, Park HA. A question answering chatbot for gastric cancer patients after curative gastrectomy: development and evaluation of user experience and performance. *Comput Inform Nurs*. 2024;42(11):829—839. https://doi.org/10.1097/CIN.0000000000001153.
14. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10(1):1. https://doi.org/10.1038/s41597-022-01899-x.
15. Xu Q, Xie W, Liao B, et al. Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: a systematic review. *J Healthc Eng*. 2023;2023:9919269. https://doi.org/10.1155/2023/9919269.
16. Meunier P-Y, Raynaud C, Guimaraes E, Gueyffier F, Letrilliart L. Barriers and facilitators to the use of clinical decision support systems in primary care: a mixed-methods systematic review. *Ann Fam Med*. 2023;21(1):57—69. https://doi.org/10.1370/afm.2908.
17. Ingraham NE, Jones EK, King S, et al. Re-aiming equity evaluation in clinical decision support: a scoping review of equity assessments in surgical decision support systems. *Ann Surg*. 2023;277(3):359—364. https://doi.org/10.1097/SLA.0000000000005661.
18. Omar M, Agbareia R, Glicksberg BS, Nadkarni GN, Klang E. Benchmarking the confidence of large language models in clinical questions. medRxiv. February 26, 2025. https://www.medrxiv.org/content/10.1101/2024.08.11.24311810v2
19. Erdat EC, Kavak EE. Benchmarking LLM chatbots' oncological knowledge with the Turkish Society of Medical Oncology's annual board examination questions. *BMC Cancer*. 2025;25(1):197. https://doi.org/10.1186/s12885-025-13596-0.
20. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Syst Rev*. 2022;18(2):e1230. https://doi.org/10.1002/cl2.1230.
21. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement : an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. https://doi.org/10.1136/bmj.n71.
22. Lockwood C, Munn Z, Porritt K. Qualitative research synthesis: methodological guidance for systematic reviewers utilizing meta-aggregation. *Int J Evid-Based Healthc*. 2015;13(3):179—187. https://doi.org/10.1097/XEB.0000000000000062.
23. Lewin S, Booth A, Glenton C, et al. Applying GRADE-CERQual to qualitative evidence synthesis findings: introduction to the series. *Implement Sci*. 2018;13(suppl 1):2. https://doi.org/10.1186/s13012-017-0688-3.
24. 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*. 2024;55:101477. https://doi.org/10.1016/j.gore.2024.101477.
25. Finch L, Broach V, Feinberg J, et al. ChatGPT compared to national guidelines for management of ovarian cancer: did ChatGPT get it right? — a Memorial Sloan Kettering Cancer Center Team Ovary study. *Gynecol Oncol*. 2024;189:75—79. https://doi.org/10.1016/j.ygyno.2024.07.007.
26. Ye Z, Zhang B, Zhang K, et al. An assessment of ChatGPT's responses to frequently asked questions about cervical and breast cancer. *BMC Womens Health*. 2024;24(1):482. https://doi.org/10.1186/s12905-024-03320-8.
27. Hermann CE, Patel JM, Boyd L, Growdon WB, Aviki E, Stasenko M. Let's chat about cervical cancer: assessing the accuracy of ChatGPT responses to cervical cancer questions. *Gynecol Oncol*. 2023;179:164—168. https://doi.org/10.1016/j.ygyno.2023.11.008.
28. Patel JM, Hermann CE, Growdon WB, Aviki E, Stasenko M. ChatGPT accurately performs genetic counseling for gynecologic cancers. *Gynecol Oncol*. 2024;183:115—119. https://doi.org/10.1016/j.ygyno.2024.04.006.
29. Yurtcu E, Ozvural S, Keyif B. Analyzing the performance of ChatGPT in answering inquiries about cervical cancer. *Int J Gynaecol Obstet*. 2025;168(2):502—507. https://doi.org/10.1002/ijgo.15861.
30. Kuerbanjiang W, Peng S, Jiamaliding Y, Yi Y. Performance evaluation of large language models in cervical cancer management based on a standardized questionnaire: comparative study. *J Med Internet Res*. 2025;27:e63626. https://doi.org/10.2196/63626.
31. Piazza D, Martorana F, Curaba A, et al. The consistency and quality of ChatGPT responses compared to clinical guidelines for ovarian cancer: A Delphi approach. *Curr Oncol*. 2024;31(5):2796—2804. https://doi.org/10.3390/curroncol31050212.
32. Levin G, Gotlieb W, Ramirez P, Meyer R, Brezinov Y. ChatGPT in a gynaecologic oncology multidisciplinary team tumour board: a feasibility study. *BJOG*. 2025;132(1):99—101. https://doi.org/10.1111/1471-0528.17929.
33. Ebner F, Hartkopf A, Veselinovic K, et al. A comparison of ChatGPT and multidisciplinary team meeting treatment recommendations in 10 consecutive cervical cancer patients. *Cureus*. 2024;16(8):e67458. https://doi.org/10.7759/cureus.67458.
34. Zou FW, Tang YF, Liu CY, Ma JA, Hu CH. Concordance study between IBM Watson for oncology and real clinical practice for cervical cancer patients in China: a retrospective analysis. *Front Genet*. 2020;11:200. https://doi.org/10.3389/fgene.2020.00200.
35. Krückel A, Brückner L, Psilopatis I, Fasching PA, Beckmann MW, Emons J. Evaluation of ChatGPT's potential in tailoring gynecological cancer therapies. *Vivo*. 2024;38(4):1649—1659. https://doi.org/10.21873/invivo.13614.
36. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. *JAMA Netw Open*. 2023;6(11):e2343689. https://doi.org/10.1001/jamanetworkopen.2023.43689.
37. Liu C, Liu X, Wu F, Xie M, Feng Y, Hu C. Using artificial intelligence (Watson for oncology) for treatment recommendations amongst Chinese patients with lung cancer: feasibility study. *J Med Internet Res*. 2018;20(9):e11087. https://doi.org/10.2196/11087.
38. Pugliese N, Wai-Sun Wong V, Schattenberg JM, et al. Accuracy, reliability, and comprehensibility of ChatGPT-generated medical responses for patients with nonalcoholic fatty liver disease. *Clin Gastroenterol Hepatol*. 2024;22(4):886—889.e5. https://doi.org/10.1016/j.cgh.2023.08.033.
39. Tan RSYC, Lin Q, Low GH, et al. Inferring cancer disease response from radiology reports using large language models with data augmentation and prompting. *J Am Med Inform Assoc*. 2023;30(10):1657—1664. https://doi.org/10.1093/jamia/ocad133.
40. Younis HA, Eisa TAE, Nasser M, et al. A systematic review and meta-analysis of artificial intelligence tools in medicine and healthcare: applications, considerations, limitations, motivation and challenges. *Diagnostics (Basel)*. 2024;14(1):109. https://doi.org/10.3390/diagnostics14010109.
41. Gulati V, Roy SG, Moawad A, et al. Transcending language barriers: can ChatGPT be the key to enhancing multilingual accessibility in health care? *J Am Coll Radiol*. 2024;21(12):1888—1895. https://doi.org/10.1016/j.jacr.2024.05.009.
42. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Med Educ*. 2023;9:e46885. https://doi.org/10.2196/46885.
43. Hincapié MA, Gallego JC, Gempeler A, Piñeros JA, Nasner D, Escobar MF. Implementation and usefulness of telemedicine during the COVID-19 pandemic: a scoping review. *J Prim Care Community Health*. 2020;11:2150132720980612. https://doi.org/10.1177/2150132720980612.
44. Monteith S, Glenn T, Geddes J, Whybrow PC, Bauer M. Commercial use of emotion artificial intelligence (AI): implications for psychiatry. *Curr Psychiatry Rep*. 2022;24(3):203—211. https://doi.org/10.1007/s11920-022-01330-7.