Journal of Multidisciplinary Healthcare

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Automating Case-Based Learning in Obstetrics and Gynecology: Validation of a Locally Deployed Large Language Model

Xueyan Hu[1,2], Xiaoju Luo[1], Yuan Zhao[1], Beibei Zheng[1]

[1]Department of Obstetrics and Gynecology, Wenzhou People's Hospital, Wenzhou, Zhejiang, People's Republic of China; [2]Department of Obstetrics and Gynecology, Chinese Medical University, Hangzhou, Zhejiang, People's Republic of China

Correspondence: Beibei Zheng, Email comeforxz_z@163.com

Background: Standardized residency training in obstetrics and gynecology (OB/GYN) relies heavily on Case-Based Learning (CBL). Yet, manually curating high-quality teaching cases remains a labor-intensive burden.

Objective: To develop and evaluate a methodological framework using a locally deployed large language model (LLM) to automate the extraction and construction of structured CBL databases.

Methods: We employed a locally deployed Qwen3-8B-Instruct model to analyze 3678 OB/GYN PubMed case abstracts, extracting primary diagnosis, core teaching points, clinical pitfalls, and difficulty levels. Two senior OB/GYN educators established a ground-truth dataset from 100 randomly selected cases to quantify extraction accuracy and AI hallucination rates. Teaching utility was assessed via a 5-point Likert scale, calculating inter-rater reliability using intraclass correlation coefficients (ICC).

Results: Our automated framework effectively categorized the 3678 cases into diverse subspecialties, with Reproductive Endocrinology and Obstetrics (28.10%) and Andrology and Male Infertility (25.20%) representing the largest cohorts. The LLM demonstrated high precision in extracting objective clinical data, achieving a 96.0% accuracy rate for primary diagnoses and 92.0% for core teaching points. Notably, the overall incidence of AI hallucinations remained low at 3.0%. In pedagogical evaluations, the model earned high scores for diagnostic accuracy (Mean = 4.48/5.0, $p < 0.05$), showing strong expert consensus (ICC = 0.92). However, the model's performance faltered when addressing subjective clinical nuances; inter-rater reliability dropped significantly regarding the utility of clinical pitfalls (ICC = 0.65) and the appropriateness of difficulty levels (ICC = 0.54).

Conclusion: This study demonstrates that LLMs offer an efficient, scalable framework for building large-scale CBL databases. While highly accurate in extracting objective facts, AI still lacks the "tacit knowledge" and empirical intuition of senior clinicians. Thus, human oversight remains indispensable to validate content and mitigate clinical risks. Future research must evaluate how these databases directly impact resident competency and learning outcomes.

Keywords: obstetrics and gynecology, OB/GYN, large language model, LLM, case-based learning, CBL, natural language processing, NLP, health informatics, knowledge extraction

Introduction

Standardized residency training in obstetrics and gynecology (OB/GYN) aims to develop independent clinical decision-making skills. However, modern OB/GYN has rapidly evolved into a highly complex, multidisciplinary field. Clinical presentations no longer exist in departmental silos; rather, they constantly intersect with andrology, clinical genetics, gynecologic oncology, and endocrinology. Traditional, single-discipline teaching methods often fail to recreate these intricate, cross-specialty clinical scenarios, creating a significant pedagogical gap.[1,2] Consequently, there is an urgent need to construct a comprehensive Case-Based Learning (CBL) database that can seamlessly integrate these fragmented knowledge domains.[3,4] While such a multidisciplinary CBL framework is essential for trainees to cultivate a holistic clinical perspective, manually curating high-quality teaching cases from exponentially growing medical literature remains a massive burden for clinical educators. Furthermore, systematically identifying "clinical pitfalls" presents

https://doi.org/10.2147/JMDH.S612064
Received: 27 March 2026
Accepted: 30 June 2026
Published: 14 July 2026

another major bottleneck. It traditionally requires a senior expert with substantial clinical intuition to identify these cognitive blind spots—which often lead to medical errors by junior physicians—before they can be formally incorporated into educational materials.[5]

The rapidly evolving large language models (LLMs) and advanced natural language processing (NLP) tools offer transformative solutions to these educational bottlenecks. Recently, the current applications of AI in obstetrics and gynecology education have garnered increasing attention, presenting both unprecedented opportunities and unique challenges.[6] Recent studies have shown that LLMs perform well in processing unstructured medical texts and accurately extracting key clinical information, such as the Generative Pre-trained Transformer (GPT).[7,8] The breakthrough in clinical reasoning benchmarks further validates the potential of artificial intelligence (AI) in an innovative approach to medical education, capable of mining thousands of published reports to identify core teaching points and hidden pitfalls. Pilot studies by Cook and Gim suggest that LLMs can rapidly integrate these narratives into interactive virtual patients and scalable, low-cost CBL modules, thereby significantly reducing faculty workload.[9,10]

Based on the above findings, this study aims to evaluate the feasibility, reliability and teaching validity of constructing a comprehensive structured CBL database using a local LLM (Qwen3-8B) to extract information from 3678 published OB/GYN case reports. To ensure data privacy and security—a critical concern in medical education—we specifically opted for a locally deployed, smaller-parameter LLM (Qwen3-8B) rather than relying on cloud-based APIs. Ultimately, this methodological validation study sought to demonstrate how AI can be effectively integrated into the clinical faculty workflow in alignment with competence-based Medical Education (CBME) requirements, providing a scalable case for modern resident standardized training.

Materials and Methods

Study Design and Data Acquisition

In this study, a retrospective text mining framework was used to construct a structured CBL database using LLM. A comprehensive literature search was conducted in PubMed database, focusing on clinical case reports in gynecology, obstetrics, reproductive medicine, and related multidisciplinary fields. To ensure methodological transparency, the precise Boolean search strategy utilized Medical Subject Headings (MeSH) and Publication Type (PT) as follows: ("Gynecology" [MeSH] OR "Obstetrics" [MeSH] OR "Gynecologic Neoplasms" [MeSH] OR "Reproductive Medicine" [MeSH] OR "Infertility" [MeSH]) AND "Case Reports" [PT].

Instead of limiting the date of publication, we selected the first 5000 cases ranked by the "best match" (association) search of PubMed. This approach preferentially retrieved highly typical and clinically representative cases, rather than extracting them purely chronologically. Automatic retrieval was performed using a custom Python script of the Biopython Entrez module (complete source code is provided in Supplementary file 1). The initial API query returned 4984 accessible records based on the relevance threshold. To ensure data quality, two authors independently screened the records, and any discrepancies were resolved through consensus discussion with a third senior author. During this rigorous data-cleaning process, a total of 1306 records were excluded for the following reasons: non-English texts or missing abstracts (n=214), non-case reports such as letters or editorials (n=685), and cases lacking clear clinical descriptions or educational value (n=407). Ultimately, the final dataset comprising 3678 high-quality abstracts was retained for automated analysis by Qwen3. Figure 1 summarizes the overall technical flow and validation process of this study.

LLM-Based Information Extraction and Structuring

To transform unstructured clinical case texts into teaching modules, we deployed Qwen3 LLM locally (specifically, the instruction-adjusted 8b parameter version Qwen3-8B-Instruct to ensure understanding of complex medical domains). To maximize the extraction consistency and reduce the hallucination of the model, we set Temperature to 0.1 and Top-P to 0.95 to strictly control the generated hyperparameters. This particular combination of parameters limits the token selection of the model to highly likely outputs, which is optimal for medical information extraction, as deterministic accuracy and factual authenticity must take precedence over creative text generation.

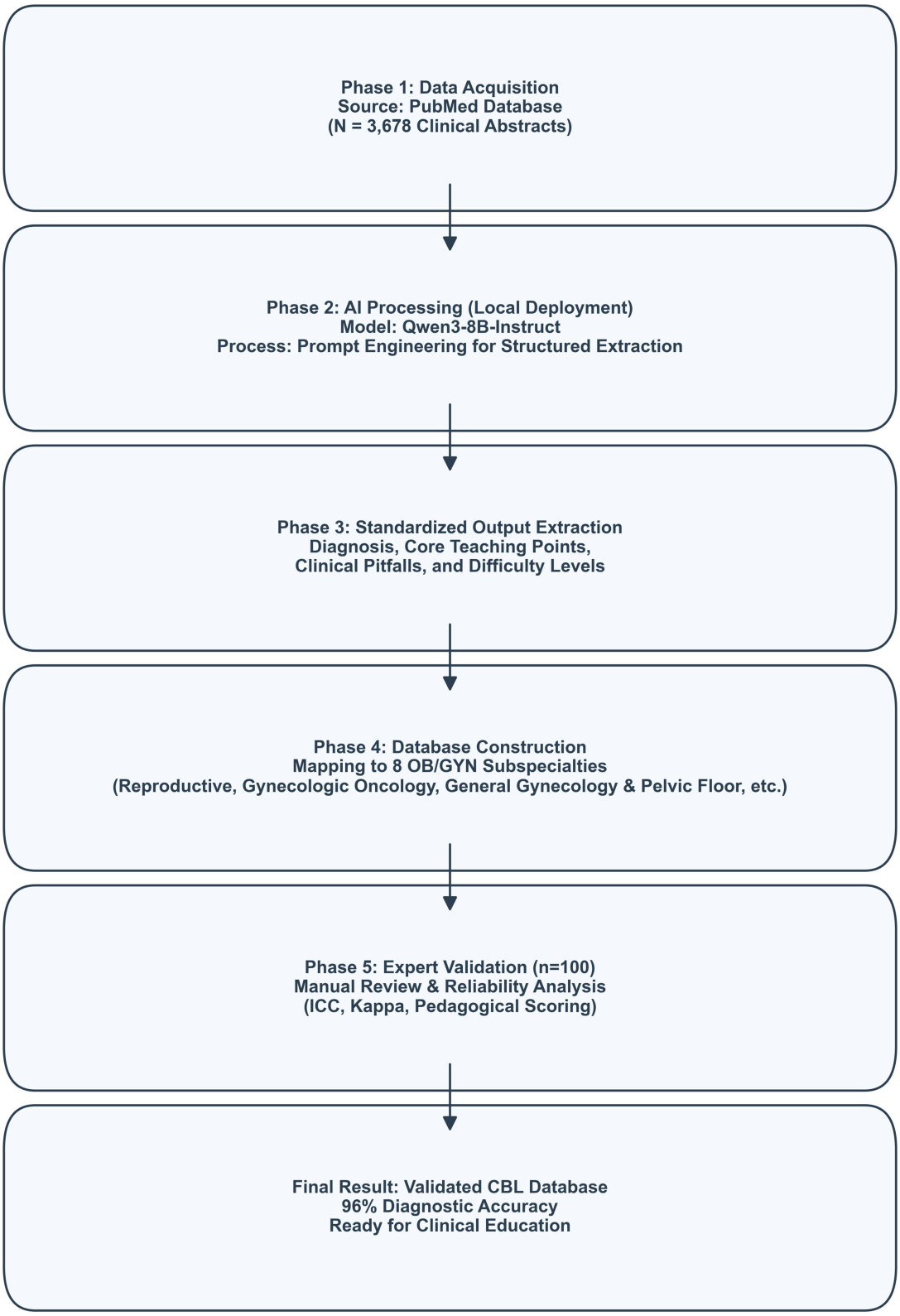

Figure 1 Methodological flowchart of the AI-driven Case-Based Learning (CBL) database construction. The process integrates automated data acquisition, structured information extraction using the Qwen3-8B model, subspecialty categorization, and rigorous expert validation to ensure pedagogical accuracy and clinical reliability.

https://doi.org/10.2147/JMDH.S612064

The model is configured with specific system prompts to act as a clinical education expert. LLM performed multi-token extraction of 4 core components on 3678 abstracts: (1) primary diagnoses, (2) teaching core points, (3) clinical pitfalls, and (4) difficulty levels.

Multidisciplinary Subspecialty Mapping

To analyze the fusion of different subspecialties, we implemented a rule-based NLP dictionary on an NLP framework designed to extract structured information and semantic annotations from clinical texts.[7,11–15] Based on the establishment of a medical surveillance framework, we developed a comprehensive bilingual medical dictionary that categorizes cases into different subspecialties. Furthermore, regular expression (Regex) matching was used to quantify the teaching topics in the extracted data.

Regarding subspecialty mapping, categories were defined based on our institutional residency rotation framework. Notably, Reproductive Endocrinology and Obstetrics were grouped together due to their overlapping focus on maternal-fetal transitions and fertility outcomes in our local training context. However, we acknowledge that these are clinically distinct domains, and this grouping strategy may obscure meaningful differences in case distributions between the two fields.

Human Validation and Quality Assessment

To rigorously evaluate the objective extraction correctness and subjective instructional utility of LLM, we performed manual validation. One hundred cases were randomly selected for human validation. The sample size calculation was performed based on Bonett's method for ICC estimation,[16] ensuring sufficient statistical power (>80%) with an alpha level of 0.05 to detect an ICC of 0.80 with a 95% confidence interval width of 0.15. To ensure reproducibility, see Supplementary file 2 for the exact system prompts used at the time of extraction.

All human evaluators involved in this study were attending physicians with over 10 years of clinical experience in OB/GYN and active roles as residency program directors or clinical educators.

Validation consisted of two stages:

1. Extraction accuracy and clinical reasoning error (hallucination) assessments: AI-generated results (diagnoses, teaching points, and subspecialties) were directly compared to the ground truth. To prevent bias, a separate panel of two senior medical educators established these gold standards. This panel operated independently of the stage II reviewers and reached a consensus on all 100 cases through discussion.

   Crucially, our system prompt explicitly instructed the LLM to infer unstated clinical pitfalls using standard medical guidelines. Therefore, we did not penalize valid, guideline-concordant inferences as errors. Instead, to ensure a rigorous pedagogical context, we quantified and redefined AI "hallucinations" and reasoning errors into two specific categories:
2. Primary Hallucinations (Guideline-Discordant Errors): Clinically incorrect inferences or recommendations. These contradict existing medical guidelines and pose potential safety risks to patients (eg, suggesting a contraindicated procedure).
3. Contextual Hallucinations (Over-extrapolations): Inferences or teaching points with general medical plausibility but lacking logical connection to the specific case narrative. This includes inappropriate over-extrapolations beyond the provided clinical context.
4. Instructional usefulness: Two independent senior experts (Reviewer A and Reviewer B) blindly rated AI-generated content on five dimensions (diagnostic accuracy, completeness of core points, subspecialty accuracy, usefulness of clinical pitfalls, and difficulty) using a 5-point Likert scale. The difficulty ratings generated by the LLM were explicitly compared with the expert-rated references.

Statistical Analysis

Statistical analyses were performed with the use of Python (version 3.12). Descriptive statistical analyses summarized baseline characteristics. To assess the correct rate of information extraction, the correct rate and hallucination rate were reported as percentages. To assess the accuracy of subspecialty mapping according to human classification, we calculated Cohen Kappa (κ). For subjective teaching evaluations (Likert scale scores), interrater reliability was quantified using

intraclass correlation coefficients (ICC, two-way mixed-effects model, absolute agreement), with 95% (CI) confidence intervals reported. Bootstrap analysis (1000 samples) was used to ensure the robustness of confidence intervals.

Ethical Considerations

Ethical approval and informed consent were not required for this study. The research did not involve human subjects, real patient encounters, or animal experiments. This study solely involved the computational text-mining and pedagogical analysis of previously published, publicly available, and de-identified case report abstracts from the PubMed database. Because this research does not meet the regulatory definition of human subjects research, it does not fall under the purview of the Institutional Review Board of Wenzhou People's Hospital, and a formal IRB exemption application was not applicable. Furthermore, the local deployment of the LLM ensured that no clinical text was transmitted to third-party cloud APIs, strictly adhering to data privacy standards.

Results

Extraction of Structured Pedagogical Components

A total of 3678 valid clinical case abstracts were processed by the LLM. As shown in Figure 2A, the most common teaching core points extracted were "differential diagnosis" (n = 1607) and "gene/genomic testing" (n = 1083). The main clinical pitfalls identified were "misdiagnosis" (n = 225) and failure to implement "individualized treatment" (n = 211), as detailed in Figure 2B. In addition, qualitative evaluation showed that the LLM could systematically synthesize these elements into structured data fields (representative cases, including an example of suboptimal AI extraction, are shown in Table 1).

Regarding difficulty levels, more than 99% of cases were classified as "hard" (52.90%, n = 1945) or "medium" (47.10%, n = 1732). Notably, only 1 case (<0.10%) was classified as "easy" (Table 2). This heavily skewed distribution likely reflects the inherent publication bias of medical case reports, which naturally tend to describe unusual, complex, or atypical presentations rather than routine clinical scenarios.

Multidisciplinary Subspecialty Mapping

Following the extraction of core components, our rule-based NLP dictionary mapped the 3678 cases into diverse subspecialties. As shown in Table 1 and Figure 3, the largest proportion of cases belonged to the Reproductive

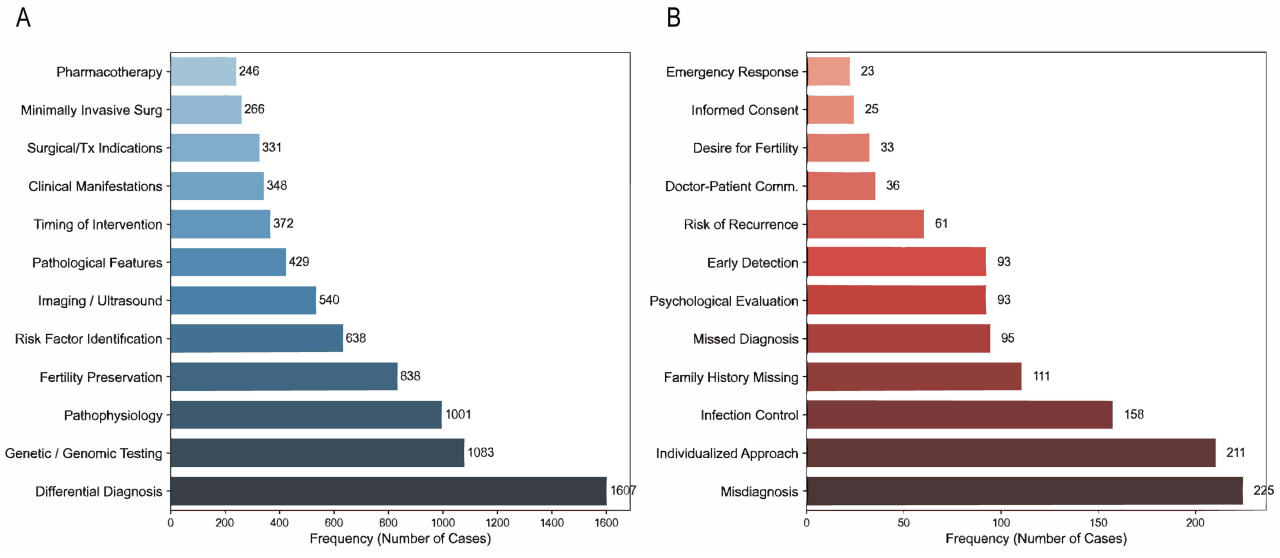

Figure 2 Frequency of pedagogical elements extracted by the LLM from the clinical case abstracts. (**A**) Distribution of the identified teaching core points, with "Differential Diagnosis" (n = 1607) and "Genetic/Genomic Testing" (n = 1083) being the most frequently extracted. (**B**) Distribution of the primary clinical pitfalls identified, highlighting "Misdiagnosis" (n = 225) and the failure to implement an "Individualized Approach" (n = 211) as the most common cognitive blind spots.

https://doi.org/10.2147/JMDH.S612064

Table 1 Representative Examples of Structured Teaching Cases Generated by Qwen3

Case Component	Example 1: Reproductive Surgery	Example 2: Gynecologic Oncology	Example 3: Rare Diseases (Suboptimal AI Extraction)
Original Article Title	The role of hysteroscopy in patients with adenomyosis and infertility: bringing out the submerged.	Gynecological tumor triplicity.	The unilateral streak gonad syndrome (Slotnick-Goldfarb syndrome).
Extracted Diagnosis	Adenomyosis complicated with infertility	Gynecological tumor triplicity	Unilateral streak gonad syndrome
Teaching Core Points	1. Differential diagnosis for repeated IVF failures. 2. Surgical indications and timing when medical therapy fails. 3. Multimodal treatment approaches	1. Surgical principles for multiple primary tumors. 2. Intraoperative decision-making and extent of surgery. 3. Differentiating borderline from invasive tumors.	1. Differential diagnosis from Turner syndrome. 2. Genetic basis and necessary testing. 3. Endocrine evaluation of hormone levels.
Clinical Pitfalls	For patients with repeated IVF failures and ultrasound evidence of adenomyotic cysts, failing to consider hysteroscopic resection may delay successful pregnancy.	When encountering an unexpected tumor during surgery, failing to comprehensively evaluate the pelvic cavity can result in missing multiple primary tumors.	Unilateral streak gonad does not equal a total loss of ovarian function. The contralateral ovary may compensate, but practitioners should beware of hormone disorders and infertility risks. Comprehensive evaluation is recommended. (Expert Note: Rated poorly as it is a generic textbook definition rather than a specific, actionable clinical warning).
Assigned Difficulty	Medium	Hard	Medium

Table 2 Baseline Characteristics of the Extracted Teaching Cases

Characteristics	Categories	Number of Cases (n=3678)	Percentage (%)
Subspecialties	Reproductive Endocrinology and Obstetrics	1035	28.10%
	Andrology and Male Infertility	927	25.20%
	Gynecologic Oncology	548	14.90%
	General Gynecology and Pelvic Floor	282	7.70%
	Family Planning and Acute Abdomen	45	1.20%
	Medical Ethics and Patient Safety	15	0.40%
	Rare Diseases and Cross-Specialties	732	19.9%
	Unclassified/Extraction Failed	94	2.60%
Difficulty Level	Hard	1945	52.90%
	Medium	1732	47.10%
	Easy/Others	1	<0.10%

Endocrinology and Obstetrics subspecialty (28.10%, n = 1035), followed by Andrology and Male Infertility (25.20%, n = 927) and Rare Diseases and Cross-Specialties (19.90%, n = 732).

Extraction Correctness and Pedagogical Inter-Rater Reliability

As shown in Table 3, the LLM demonstrated robust objective extraction correctness compared to human experts. The accuracy of primary diagnostic extraction was 96.0%, and the accuracy of the subspecialty mapping algorithm against manual classification was 94.0% (Cohen's $\kappa = 0.88$).

Importantly, the overall incidence of AI hallucinations remained at a very low 3.0%. When broken down by taxonomy, primary hallucinations (guideline-discordant clinical errors) accounted for 1.0%, while contextual hallucinations (plausible but over-extrapolated points not explicitly supported by the original text) accounted for 2.0%.

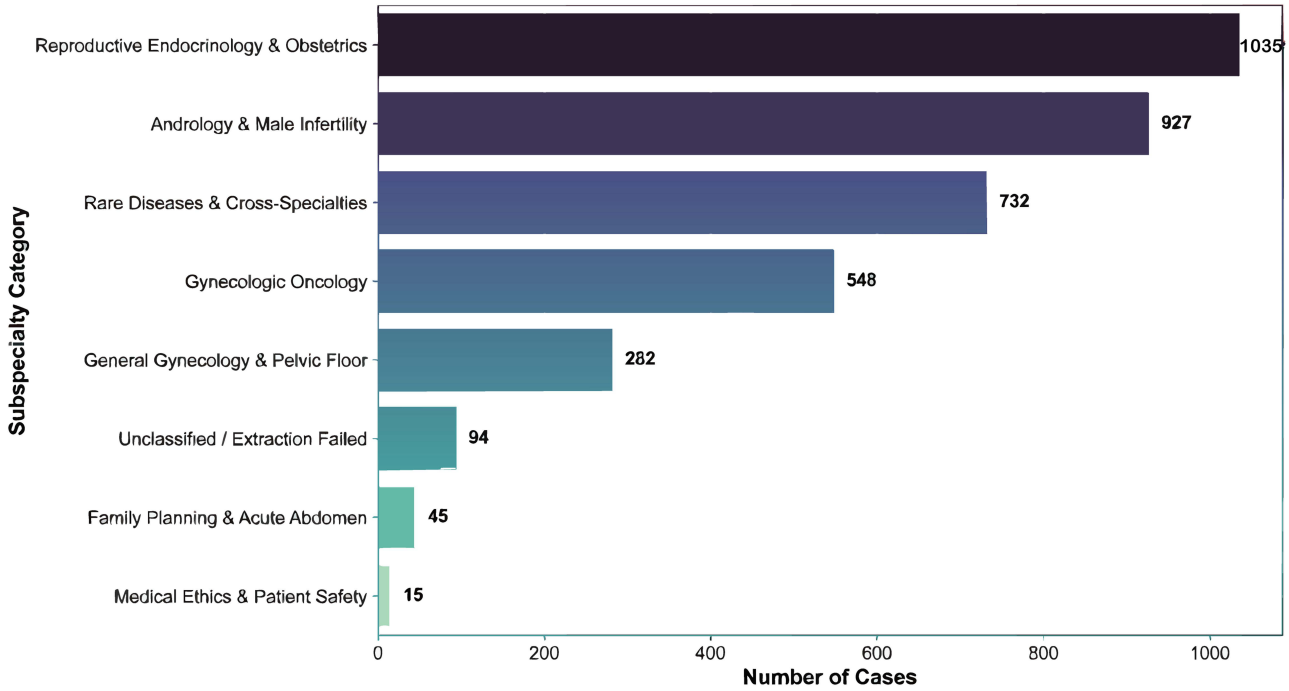

Figure 3 Distribution of the extracted case reports across multidisciplinary subspecialty categories. The bar chart illustrates the absolute frequency of cases (n = 3678) mapped to each domain, highlighting the predominance of Reproductive Endocrinology and Obstetrics (n = 1035) and Andrology and Male Infertility (n = 927).

However, when assessing the utility of subjective clinical pitfalls, the factual error rate representing AI hallucinations increased to 6.0%. This error predominantly occurred because the LLM struggled to accurately infer implicit clinical context that was not explicitly stated in the original text, occasionally leading to overdefensive or clinically inappropriate recommendations. Consequently, inter-rater reliability dropped significantly for these subjective nuances, resulting in an ICC of 0.65 [95% CI: 0.50–0.76] for clinical pitfalls and an ICC of 0.54 [95% CI: 0.35–0.68] for the appropriateness of difficulty levels.

Discussion

Principal Findings

This study demonstrates the capacity of a locally deployed LLM (Qwen3-8B-Instruct) to systematically curate an extensive Case-Based Learning (CBL) database comprising 3678 clinical reports. The automated extraction significantly alleviates the cognitive and temporal burden placed on clinical faculty. Our framework achieved robust objective correctness, with a 96.0% accuracy rate for diagnostic extraction and a 94.0% accuracy for multidisciplinary subspecialty

Table 3 Extraction Accuracy, Hallucination Rate, and Pedagogical Inter-Rater Reliability (n=100 Subset)

Evaluation Metrics	Objective Extraction Accuracy (%) vs Ground Truth	Factual Error/ Hallucination Rate (%)	Pedagogical Mean Score (1–5 scale)	Inter-Rater Reliability (ICC/κ) [95% CI]
Diagnostic Accuracy	96.00%	1.00%	4.48	ICC = 0.92 [0.87–0.95]
Completeness of Core Points	92.00%	3.00%	4.36	ICC = 0.89 [0.84–0.93]
Accuracy of Subspecialty	94.00%	2.00%	4.25	κ = 0.88 [0.82–0.94]
Utility of Clinical Pitfalls	N/A (Subjective inference)	6.0%*	4.12	ICC = 0.65 [0.50–0.76]
Appropriateness of Difficulty	81.0% (vs Expert reference)	N/A	4.44	ICC = 0.54 [0.35–0.68]

Notes: *For clinical pitfalls, the factual error rate represents the frequency of AI-generated "hallucinations" (clinically inappropriate or unsafe recommendations). Bootstrap validation (1000 resamples) was applied for CI estimations.
Abbreviations: ICC, Intraclass Correlation Coefficient; CI, Confidence Interval.

https://doi.org/10.2147/JMDH.S612064

mapping (Cohen's κ = 0.88). Importantly, the overall incidence of true hallucinations remained remarkably low at 3.0% when extracting objective teaching points. Furthermore, our subspecialty mapping highlighted an increasing necessity for deep multidisciplinary integration within modern OB/GYN, as the published case report literature suggests that disciplines like andrology, clinical genetics, and oncology now constitute a substantial portion of the relevant clinical content. Traditionally, medical education often suffers from "siloed" teaching, where human educators may naturally focus on nuances within their specific subspecialty.[17] By providing AI-facilitated identification of cross-disciplinary case content, the LLM effectively categorizes these existing intersections from the literature. Rather than acting as a standalone creator of integration, the model extracts these structured fields independently, thereby helping to breaks down departmental boundaries to support holistic OB/GYN training. This multi-domain integration is critical for modern competence-based medical education, as it can foster a more comprehensive view among junior physicians, equipping them to navigate complex, multi-system diseases.

Comparison with Prior Work

Our findings align with contemporary evidence highlighting the proficiency of LLMs in processing complex clinical narratives for educational applications, building upon earlier studies demonstrating their potential in medical education.[7,8,10] Furthermore, our findings are consistent with the concept of a "tacit knowledge" gap between AI and human faculty. While the LLM excelled at explicit textbook-level knowledge extraction, we observed significant performance drops when evaluating subjective clinical nuances—a phenomenon echoed in previous studies comparing AI and human clinical reasoning ratings.[18,19] For instance, when the model attempted to extrapolate clinical pitfalls, the factual error rate representing clinically inappropriate recommendations increased to 6.0%, dropping the inter-rater reliability (ICC = 0.65) for this metric. An illustrative example occurred in an emergency obstetric case during the COVID-19 pandemic (PMID: 33218391), where the AI made an overdefensive and clinically inappropriate recommendation for proactive cesarean delivery, contradicting existing ACOG guidelines. This demonstrates that while current generative AI effectively extracts explicit information, it struggles to reliably simulate the empirically-driven intuition and situational risk assessment inherent to senior clinicians.

Practical Implications

From an implementation perspective, we deliberately opted for a local deployment of the 8B-parameter model rather than relying on larger cloud-based APIs. This approach provides a secure design advantage by offering a scalable and privacy-preserving solution, eliminating the data breach risks associated with transmitting clinical narratives to external servers—a critical consideration for medical informatics. To translate these technical findings into clinical curricula, we propose a "human-in-the-loop" constructivist workflow (Figure 4) for future implementation. When educators query the database for multidisciplinary topics, the LLM produces structured data that could form the basis for CBL module development. Human faculty must subsequently review these extracted elements to filter out "primary hallucinations" and refine tacit clinical pitfalls before synthesizing them into ready-to-use modules for residents. This proposed paradigm maximizes

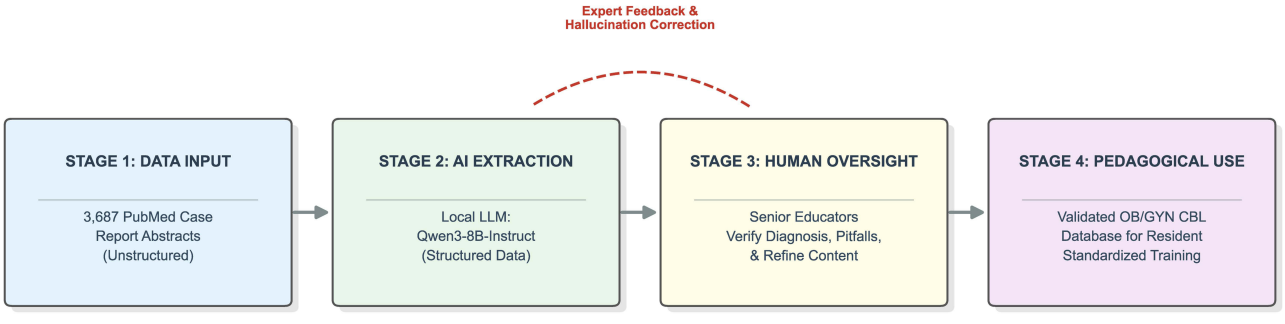

Figure 4 The Proposed AI-Assisted Human-in-the-Loop Teaching Workflow. This schema illustrates the four-stage evolution of the CBL database. It highlights the critical "human-in-the-loop" phase (indicated by the red curved arrow), where senior educators provide direct feedback and correct potential AI hallucinations, transforming raw AI extractions into validated pedagogical tools.

AI's efficiency while preserving the critical mentoring role of clinical teachers, aligning with evolving clinician educator milestones.[20] Furthermore, it must be emphasized that obstetrics and gynecology is a specialty characterized by ss, emergency scenarios requiring rapid decision-making, and maternal-fetal ethical conflicts. Therefore, AI hallucinations in this field may lead to much more serious consequences compared to general internal medicine, making stringent human oversight an absolute necessity.

Limitations

Several limitations warrant consideration. First, regarding data acquisition, the use of the PubMed "Best Match" algorithm may introduce significant selection bias, and over 99% of the extracted cases were categorized as "medium" or "hard". This extreme tilt reflects the inherent publication bias of peer-reviewed literature, which naturally prioritizes complex or atypical scenarios. While highly advantageous for senior residents under the Competency-Based Medical Education (CBME) framework,[21] pushing them toward expert-level clinical reasoning,[22–24] it may overwhelm junior medical students.[25]

Second, concerning data depth, our analysis was performed exclusively at the abstract level rather than utilizing full-text case reports. Therefore, the extracted structured content should not be overinterpreted in terms of full "educational completeness".

Third, regarding AI hallucinations and inherent biases, although we reported overall numerical hallucination rates, we did not perform a detailed sub-analysis to identify which subspecialty areas, clinical scenarios, or prompt structures were most vulnerable to errors. Additionally, the risk of subtle factual inaccuracies persists; future deployments must remain vigilant regarding systematic or demographic biases inadvertently perpetuated by LLMs in clinical narratives.[26,27]

Fourth, our expert validation methodology has constraints in external validity. Evaluating only a subset of 100 out of 3678 cases limits the broader generalizability of the findings. Moreover, the validation panel consisted entirely of OB/GYN specialists from a single center. It remains to be discussed whether similar performance could be achieved across different countries and educational systems. Future iterations would also benefit from including General Practice (GP) educators to assess the utility of these cases for primary care screening.

Fifth, regarding long-term model stability, LLMs rapidly evolve over time. The locally deployed Qwen3-8B model used in this study will likely be updated, meaning the long-term stability and exact reproducibility of these extraction outputs remain uncertain.

Finally, concerning pedagogical impact, this remains a methodological validation study; we have yet to assess the database's direct impact on resident learning outcomes or competency scores via interventional trials.

Conclusions

This study demonstrates that locally deployed LLMs offer an efficient, scalable, and secure methodological framework facilitating the construction of large-scale CBL databases. Our validation confirms that AI can achieve high extraction accuracy for objective core knowledge while maintaining low hallucination rates. However, these findings primarily establish the pedagogical acceptability of the AI-generated content rather than its definitive educational effectiveness. Consequently, oversight by human experts remains an indispensable safeguard for screening empirically driven tacit knowledge, mitigating clinical risks, and ensuring patient safety in clinical education. Future interventional research must evaluate how integrating these AI-generated modules into standardized residency training directly impacts physician competency and diagnostic performance.

Data Sharing Statement

The datasets generated and analyzed during the current study, including the structured Case-Based Learning (CBL) database extracted by the LLM, are available from the corresponding author upon reasonable request. The original raw clinical case abstracts analyzed in this study are publicly available in the PubMed database.

Acknowledgments

The authors would like to thank the clinical educators and fellowship directors who participated in the blinded human validation process for their invaluable insights.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

The authors declare that no funding or research grants were received for this study.

Disclosure

The authors declare that they have no competing interests in this work.

References

1. Weiss TG, Rentea RM. Simulation training and skill assessment in obstetrics and gynecology. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing; 2025.
2. Wittek A, Strizek B, Recker F. Innovations in ultrasound training in obstetrics. *Arch Gynecol Obstet*. 2025;311(3):871–880. doi:10.1007/s00404-024-07777-8
3. Florek AG, Dellavalle RP. Case reports in medical education: a platform for training medical students, residents, and fellows in scientific writing and critical thinking. *J Med Case Rep*. 2016;10:86. doi:10.1186/s13256-016-0851-5
4. Jackson D, Cleary M, Hickman L. Case reports as a resource for teaching and learning. *Clin Case Rep*. 2014;2(5):163–164. doi:10.1002/ccr3.172
5. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad Med*. 2003;78(8):775–780. doi:10.1097/00001888-200308000-00003
6. Torumtay Aliç SB. The role of artificial intelligence in obstetrics and gynecology: innovations, challenges, and opportunities explored through a bibliometric analysis. *Int J Gynaecol Obstet*. 2026;174:210–218. doi:10.1002/ijgo.70797
7. Kim MS, Chung P, Aghaeepour N, Kim N. Information extraction from clinical texts with generative pre-trained transformer models. *Int J Med Sci*. 2025;22(5):1015–1028. doi:10.7150/ijms.103332
8. Sciannameo V, Pagliari DJ, Urru S, et al. Information extraction from medical case reports using OpenAI InstructGPT. *Comput Methods Programs Biomed*. 2024;255:108326. doi:10.1016/j.cmpb.2024.108326
9. Cook DA. Creating virtual patients using large language models: scalable, global, and low cost. *Med Teach*. 2025;47(1):40–42. doi:10.1080/0142159x.2024.2376879
10. Gim H, Cook B, Le J, et al. Large language model-supported interactive case-based learning: a pilot study. *Intern Med J*. 2025;55(5):852–855. doi:10.1111/imj.70030
11. Campillos-Llanos L, Valverde-Mateos A, Capllonch-Carrión A. Hybrid natural language processing tool for semantic annotation of medical texts in Spanish. *BMC Bioinformatics*. 2025;26(1):7. doi:10.1186/s12859-024-05949-6
12. Gao C, Gheihman G, Kaplan T, et al. Education research: creating online interactive case-based learning experiences from educational case reports with large language models: a feasibility study. *Neurol Educ*. 2025;4(4):e200250. doi:10.1212/ne9.0000000000200250
13. Herman Bernardim Andrade G, Yada S, Aramaki E. Is boundary annotation necessary? Evaluating boundary-free approaches to improve clinical named entity annotation efficiency: case study. *JMIR Med Inform*. 2024;12:e59680. doi:10.2196/59680
14. Shibata D, Shinohara E, Shimamoto K, Kawazoe Y. Towards structuring clinical texts: joint entity and relation extraction from Japanese Case Report Corpus. *Stud Health Technol Inform*. 2024;310:559–563. doi:10.3233/shti231027
15. Shinohara E, Shibata D, Kawazoe Y. Development of comprehensive annotation criteria for patients' states from clinical texts. *J Biomed Inform*. 2022;134:104200. doi:10.1016/j.jbi.2022.104200
16. Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. *Stat Med*. 2002;21(9):1331–1335. doi:10.1002/sim.1108
17. Frenk J, Chen L, Bhutta ZA, et al. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. *Lancet*. 2010;376(9756):1923–1958. doi:10.1016/s0140-6736(10)61854-5
18. Rajan A, Alexander SM, Shenvi CL. Can AI grade like a professor? Comparing artificial intelligence and faculty scoring of medical student short-answer clinical reasoning exams. *Adv Health Sci Educ Theory Pract*. 2025;31:607–617. doi:10.1007/s10459-025-10462-3
19. Waldock WJ, Lam G, Baptista A, Walls R, Sam AH. Which curriculum components do medical students find most helpful for evaluating AI outputs? *BMC Med Educ*. 2025;25(1):195. doi:10.1186/s12909-025-06735-5
20. Mahan JD, Kaczmarczyk JM, Miller Juve AK, et al. Clinician educator milestones: assessing and improving educators' skills. *Acad Med*. 2024;99(6):592–598. doi:10.1097/acm.0000000000005684
21. Agrawal A, Sharma A, Sharma A, Agrawal C. Challenges faced by medical faculty in implementation of competency-based medical education and lessons learned. *J Educ Health Promot*. 2024;13:345. doi:10.4103/jehp.jehp_892_23

22. Anandan C, Anderson NC, Barbour K, et al. Enhancing neurologic clinical competency-implementing competency-based medical education and workplace-based assessments in a neurology clerkship. *Semin Neurol*. 2025;46:257–262. doi:10.1055/a-2762-9535
23. Anteby E. [Competence based medical education]. *Harefuah*. 2025;164(4):260–264.
24. Bakunda L, Crooks R, Johnson N, et al. Redefining professionalism to improve health equity in competency based medical education (CBME): a qualitative study. *MedEdPublish*. 2024;14:237. doi:10.12688/mep.20489.1
25. Abdullahi T, Singh R, Eickhoff C. Learning to make rare and complex diagnoses with generative AI assistance: qualitative study of popular large language models. *JMIR Med Educ*. 2024;10:e51391. doi:10.2196/51391
26. Menz BD, Kuderer NM, Chin-Yee B, et al. Gender representation of health care professionals in large language model-generated stories. *JAMA Netw Open*. 2024;7(9):e2434997. doi:10.1001/jamanetworkopen.2024.34997
27. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health*. 2024;6(1):e12–e22. doi:10.1016/s2589-7500(23)00225-x