Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Precision Education in Orthopedic Surgery: Impact of an AI-Driven RAG Model on Perioperative Nursing Students

Sedighe Hannani

Sedighe.hannani.iums@gmail.com

Iran University of Medical Sciences

Bahador Pourdel

Iran University of Medical Sciences

Fardin Amiri

Iran University of Medical Sciences

Zeinab Barzegar

Iran University of Medical Sciences

Namaamali Azaadi

Iran University of Medical Sciences

Research Article

Keywords: Artificial Intelligence, Retrieval-Augmented Generation, Perioperative Nursing, Precision Education

Posted Date: March 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8688674/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Introduction: The complexity of orthopedic surgery causes the minds of perioperative nursing students to be stressed, and thus, they need to be better educated. The transformative potential of generative AI is real, and general-purpose Large Language Models (LLMs) are not domain-specific and hallucinate. The effect of an orthopedic surgical technology course on learning outcomes and student satisfaction will be analysed with the help of a domain-specific Retrieval-Augmented Generation (RAG) model, which is based on authoritative literature on surgery.

Methods: 50 perioperative nursing students in two Tehran medical universities were selected to take part in a quasi-experimental non-equivalent pretest-posttest controlled group. The control group (n=25) was given conventional teaching, and the intervention group (n=25) was given RAG-based intelligent teaching. The Education Satisfaction Questionnaire and the Orthopedic Surgical Technology Knowledge Test were verified. Mann-Whitney U and Wilcoxon tests were done by SPSS.

Results: The difference in learning was significant in the intervention group, and the effect size was greater (Mean Gain: 15.72 vs. 12.64, $P < 0.001$, $r = 0.77$). Team RAG-based system had a moderate-high user satisfaction (Mean: 72.48+7.56), which means that the technology was used successfully and positively influenced the user experiences.

Discussion: The domain-specific RAG model might minimize cognitive load and grounded retrieval hallucinations as an intelligent tutoring support. These findings indicate that precise surgical teaching devices can enhance cognitive education and change education. This knowledge should be taken long-term through research and clinical use.

1. Introduction

Surgical procedures, as a crucial aspect of healthcare quality worldwide, have always relied on the competence and proficiency of surgical teams (1). The recent decades have seen the growing complexity of orthopedic surgery brought about by the introduction of sophisticated prosthetics, an enormous number of instruments, and new minimally invasive procedures that have turned the operating room into a high-stakes game requiring the highest level of precision (2). In this regard, patient safety can suffer irreparable consequences due to human errors stemming from insufficient theoretical knowledge and clinical skills (3). This has led modern paradigms of medical education to shift the emphasis from traditional one-size-fits-all methods towards Precision Education (4). This strategy is based on factual information, individualized learning, and the use of smart technologies to reduce cognitive errors and improve clinical outcomes (5, 6). The need to develop a qualified workforce that can make accurate evidence-based decisions in critical situations has made it a strategic requirement to incorporate the use of sophisticated educational technologies in education programs (7).

The perioperative nursing students face complex cognitive tasks in this intricate ecosystem; hundreds of specialized orthopedic tools, understanding the order of the surgery, and sterility control can become a

heavy Cognitive Load on the working memory (8, 9). Conventional pedagogical approaches, which are mostly based on the passive delivery of knowledge, may well fail to handle this amount of information and convert it into applied knowledge (10). To address this issue, Generative Artificial Intelligence models, especially Large Language Models (LLMs), offer opportunities to transform education (11). Nonetheless, general-purpose model implementation in specialized fields is limited to issues like "Hallucination" and incompatibility with authoritative scientific literature (12). One of the new ways to beat this challenge is the use of Retrieval-Augmented Generation (RAG) models (13). The RAG architecture links the language model to a domain-specific, authoritative body of knowledge to produce correct, grounded, and citable answers that can improve both important indicators of educational quality: Learning Outcomes and Student Satisfaction (14).

A review of scientific sources indicates that the integration of digital technology, such as Virtual Reality (VR) and intelligent simulators in surgical training, has had a positive impact on technical competence and attention to learners (15). As per the previous studies, AI-based tools, due to the possibility of offering instant feedback and minimizing the extraneous cognitive load, can shorten the learning curve (16). However, it was found that student confidence in them is weak, and the use of generic language models, which are characterized by knowledge cutoffs, may result in obtaining out-of-date or inaccurate information (17). Despite the abundance of research literature on the topic of the overall use of AI in medicine (18), there is little empirical material on the effectiveness of RAG models specifically adjusted to high-stakes and accurate training, e.g., orthopedic surgical technology, and the drawbacks of conventional digital tools are still present (19).

In spite of the fact that AI has been increasing exponentially in medical education, there is a glaring research gap concerning the methodical study of the effects of domain-specific RAG models on perioperative nursing education. Scientifically speaking, it is not yet clear enough how the substitution of the parametric knowledge of general models with the knowledge obtained from the special orthopedic sources can produce a considerable difference in cognitive learning results and user satisfaction. Moreover, in the indigenous background of Iranian higher education (which is characterized by limited access to standardized clinical settings, a low ratio of clinical instructors to students, and an urgent need for cost-effective and scalable educational solutions), the effectiveness of such an intelligent system is two-fold. The lack of research that has tested this technology in a real-life surgical technology teaching environment highlights the importance of carrying out this research.

Thus, the purpose of the given research is to explore how a domain-specific RAG model can influence perioperative nursing students' learning outcomes and satisfaction in the Orthopedic Surgical Technology course. It is hoped that the results of this study will not only give solid empirical data indicating the effectiveness of AI-based precision education but also a new paradigm of designing intelligent tutoring systems that, by ensuring scientific quality and reducing errors in the teaching process, will eventually improve the quality of surgical care in the future.

2. Methods

The study was planned and carried out as a quasi-experimental research with a non-equivalent pretest-posttest control group design in order to explore the efficacy of an intelligent educational model.

2.1. Study Design and Setting

This research was carried out in the first semester of the 2025–2026 academic year in two different academic environments in Tehran. The allocation of groups was done according to the educational setting and, therefore, students of the School of Allied Medical Sciences at the Iran University of Medical Sciences were put in the intervention group (RAG-based education), and students of the School of Allied Medical Sciences at the AJA University of Medical Sciences in the control group (conventional education), to avoid contamination bias. Both facilities followed the standard curricula that were certified by the Ministry of Health as the course of Orthopedic Surgical Technology.

The independent variable in this design was the RAG model-based educational intervention, and the dependent variables were cognitive learning and educational satisfaction. Also, an initial difference was controlled for by including a covariate, the pretest score.

2.2. Population, Sampling, and Participants

The study population included all fifth-semester undergraduate perioperative nursing students enrolled in the Orthopedic Surgical Technology course in the field of surgical technology. Non-probability convenience sampling was used.

G Power software was used to calculate the sample size, with a test power of 0.80, a significance level of 0.05, and a medium effect size (d = 0.6). The estimated sample size was 33 participants per group, accounting for a 10% attrition rate. Inclusion criteria were as follows: fifth-semester student, taking the course in question, and filling out the informed consent form. The exclusion criteria were not attending at least three sessions, not filling in questionnaires, and sharing the tool of intervention (in the case of the control group). Finally, the final analysis involved 50 students (25 per group) who went through all the stages.

2.3. Data Collection Instruments and Psychometrics

Two researcher-developed data collection instruments (see Supplementary File 1 for the full English versions of the instruments), which were validated in a rigorous process, were used to collect data:

- **Orthopedic Surgical Technology Knowledge Test**: The purpose of this test was to measure learning, and a 40-item multiple-choice test was created on the topics of anatomy, pathophysiology, instrumentation, surgical techniques, and intraoperative care. Scores ranged from 0 to 40.
- **Educational Satisfaction Questionnaire**: To evaluate the user experience in the intervention group, a 20-item questionnaire was created on a 5-point Likert scale (score range: 20–100).

Validity and Reliability

The 10 senior students revised and refined face validity. A panel of 10 experts was used to assess the content validity, according to Lawshe, in a table, and the approach suggested by Waltz and Bausell, all the items with Content Validity Ratio (CVR) less than 0.62 and Content Validity Index (CVI) less than 0.79 will be eliminated or revised. The instrument reliability was tested through a pilot study on 35 students, and the alpha coefficients of Cronbach were more than 0.7 in both tools, which is a good indication of internal consistency.

2.4. RAG System Architecture and Intervention Protocol

The educational course was implemented in 16 weeks after the administration of the pretest in the two groups. After the fourth week, students in both groups were introduced to 11 complex clinical situations. The control group examined the situations with the help of conventional tools (textbooks and articles), and the intervention group operated only with the help of a developed intelligent System. A posttest and the satisfaction survey (intervention group specific) were conducted at the end of the course.

Technical Architecture

The System used is a particular chatbot based on the Retrieval-Augmented Generation (RAG) architecture, implemented in Python. Figure 1 gives the schematic diagram of this architecture with the flow of data processing. It is made up of four layers of processing:

- **Data Ingestion**: The data of reference sources (such as the Operating Room Technique by Berry and Kohn, Surgical Technology for the Surgical Technologist, and Alexander's Care of the Patient in Surgery) were scrubbed and separated into 900-character blocks with a 100-character window using a Sliding Window strategy.
- **Indexing**: BAAI/bge-large-en embedding model was used to vectorize text chunks and index them with the FAISS index to enable them to be quickly retrieved.
- **Bilingual Retrieval**: In order to overcome the language issue, user queries in the Persian language were initially translated into specialized English with the help of the Gemini 2.5 model. After that, the most relevant text chunks (Context) were extracted by means of the vector search.
- **Generation**: Ultimately, the DeepSeek V3 language model, fed with the query provided by the user, the texts retrieved, and under strict System Prompts to avoid hallucinations, generated the final answer in Persian. The user interface was built and implemented on the Streamlit platform to make it easy to access and use. Figure 2 shows a screenshot of the chatbot interface in an active session with the query typed by the user in Persian, the formatted response the model produces, and additional clinical reasoning that has been given to the student.

Validation of Translation and Terminological Accuracy

The quality of the output of the system was also strictly tested by the same panel of 10 subject matter experts (orthopedic surgeons and senior surgical technologists) who validated the study instruments to counter the risk of semantic loss in the bilingual processing (Persian-English-Persian). The experts have

revised a random sample of 50 system-generated responses to complex clinical questions before the intervention. They particularly evaluated the precision of translated specialized terminology (e.g., particular retractors, hemostatic agents, and prosthetic elements) and clinical thinking. The panel established that there was a high level of clinical accuracy and terminological consistency of the dual-translation architecture and that no serious hallucinations or mistranslations of the critical surgical setting were observed. Nevertheless, the frequency of hallucinations was not quantitatively assessed in the intervention, and only the expert analysis of the system output before the intervention was conducted.

2.5. Data Analysis Method

SPSS version 26 was used to analyze the data, and P was set at $P < 0.05$. The Shapiro-Wilk test was used to determine whether the continuous variables were normally distributed. Since the primary outcome variables have been distributed in a manner that could not allow the assumption of normality, and the sample size was rather small, the non-parametric statistical techniques were used exclusively.

The Wilcoxon signed-rank test was used to determine within-group changes between pre-test and post-test scores. The Mann-Whitney U test was applied to provide between-group comparisons. Mean ± standard deviation was reported as a descriptive statistic of normally distributed variables, and median and interquartile range as descriptive statistics of non normally distributed variables.

3. Results

Data from 50 perioperative nursing students (25 in the intervention group and 25 in the control group) were used in this study. Considering the outcomes of the Shapiro-Wilk test, which showed that there was a non-conformity to normal distribution in the significant variables (pre-test scores and age) ($P < 0.05$), inferential tests were conducted using non-parametric tests. The level of significance of all tests was set to 0.05.

3.1. Assessment of Psychometric Properties of Instruments

The validity and reliability indices of the measurement instruments were investigated before the analysis of the main data. Table 1 shows that all instruments have acceptable scores on the Content Validity Ratio (CVR) and Content Validity Index (CVI). Moreover, the Cronbach alpha coefficients (more than 0.79) obtained after the calculations are high, which indicates high internal consistency and reliability of the instruments used in data collection.

Table 1
Psychometric Properties of Research Instruments

Instrument Name	Number of Items	Mean CVR	Mean CVI	Reliability (Cronbach's Alpha)	Final Status
Surgical Knowledge Test	40	0.815	0.892	0.790	Confirmed
Satisfaction Questionnaire	20	0.850	0.920	0.819	Confirmed

3.2. Demographic Characteristics and Baseline Homogeneity

The analysis of demographic and academic traits at study entry showed that there were no statistically significant differences between the two groups (Table 2). Age, gender, and the average GPA of the last semester were the same in both groups. Also, the comparison of the pre-test score with the help of the Mann-Whitney U test (P = 0.088) proved that the participants of both groups had an equal level of background knowledge; this conclusion excludes the influence of initial confounding factors on the final data.

Table 2
Comparison of Demographic and Baseline Variables Between Groups (Homogeneity Check)

Variable	Group	Mean ± SD / N (%)	Test Statistic	P-value	Effect Size (r or ϕ)	Result
Age (Years)	Intervention	20.84 ± 0.85	Z = -0.269	0.788	0.04 (Negligible)	Homogeneous
	Control	20.88 ± 0.72				
Total GPA	Intervention	17.83 ± 0.47	Z = -0.825	0.410	0.12 (Small)	Homogeneous
	Control	17.68 ± 0.43				
Pre-test Score	Intervention	10.84 ± 1.31	Z = -1.704	0.088	0.24 (Small)	Homogeneous
	Control	10.16 ± 1.28				
Gender	Intervention	Male: 11 (44%) / Female: 14 (56%)	Chi-Square = 0.321	0.571	0.08 (Negligible)	Homogeneous
	Control	Male: 13 (52%) / Female: 12 (48%)				

3.3. Analysis of Learning Outcomes

The analysis of the impact of the educational intervention on the scores of students in terms of specialized knowledge was carried out on two scales: intra-group changes and inter-group differences.

3.3.1. Temporal Changes in Scores (Intra-group)

The Wilcoxon test was used to compare the pre-test and post-test scores to investigate the trend of learning in each group (Table 3). The findings showed that there was a great improvement in both the intervention and control group post-test scores ($P < 0.001$). This observation indicates that, despite the type of educational approach that was used, the course resulted in knowledge acquisition among the students.

Table 3
Comparison of Mean Pre-test and Post-test Scores Within Groups (Wilcoxon Test)

Group	Test Phase	Mean ± SD	Direction of Change (Pos/Neg)	Z Statistic	P-value	Effect Size (r or ϕ)
Control (Traditional Education)	Pre-test	10.16 ± 1.28	25 Pos / 0 Neg	-4.410	< 0.001*	0.62 (Large)
	Post-test	22.80 ± 1.63				
Intervention (RAG Model)	Pre-test	10.84 ± 1.31	25 Pos / 0 Neg	-4.401	< 0.001*	0.62 (Large)
	Post-test	26.56 ± 1.91				

3.3.2. Comparison of Educational Method Effectiveness (Inter-group)

The academic achievement (Score Gain), which was computed as the difference in the post-test and pre-test scores, was used to compare the two groups to determine the efficiency of the proposed model. Table 4 and Fig. 3 show the statistically significant difference in support of the intervention group ($P < 0.001$). In particular, the mean learning outcome among the students who used the RAG model (15.72) was significantly higher than that of the students in the control group of education (12.64).

Table 4
Comparison of "Learning Gain" (Gain Score) Between Groups (Mann-Whitney U Test)

Variable	Group	Mean Gain (Score)	Sum of Ranks	Z Statistic	P-value	Effect Size (r or ϕ)
Learning Gain (Post-test - Pre-test)	Control	+ 12.64	359.00	-5.466	< 0.001*	0.77 (Very Large)
	Intervention	+ 15.72	916.00			

3.4. Learner Satisfaction Status

Table 5 shows the level of user satisfaction in the intervention group (N = 25) in terms of interaction with the intelligent educational system. The mean score of satisfaction was 72.48 (SD = 7.56) on a scale of

20–100. This score represents a moderate to high level of satisfaction, Reporting a fairly good user experience among students who were exposed to the RAG-based system.

Table 5
Descriptive Statistics of Satisfaction Scores in the Intervention Group

Variable	Count	Min Score	Max Score	Mean	SD	Qualitative Interpretation
Satisfaction Score	25	56	89	72.48	7.56	Moderate to High

4. Discussion

The current research was designed as a quasi-experimental study that tried to investigate the effectiveness of using the Retrieval-Augmented Generation (RAG) architecture in the delivery of the specialized course of Orthopedic Surgical Technology. The findings of data analysis outline an exciting future for the evidence-based Artificial Intelligence application in the education of surgical sciences. The utilization of this language model was found to result in a significant improvement of the cognitive learning indicators and achievement of a desirable level of satisfaction among perioperative nursing students through inferential findings. These findings are then accurately expounded in the ensuing sections against a theoretical background, previous empirical studies, and critical reviews.

4.1. Methodological Validation and Group Homogeneity

The introduction of interventional findings requires the initial equivalence of the research units in order to interpret them correctly. Statistical tests showed that the confounding variables (age, gender, and academic background (GPA)) distribution was homogenous in both groups, and no significant difference was found in the pre-test scores ($P > 0.05$). This preliminary correspondence ensures internal validity of the study, which indicates that the differences between the post-test results can probably be attributed to the educational intervention and not to individual differences at baseline. This observation is in line with the methodological principles that have been developed in controlled trials that insist on the need to control confounding variables to make causal inferences.

4.2. Analysis of Intervention Effectiveness on Cognitive Learning Outcomes

The main results of the study show that there is statistically significantly greater success in the final assessments of the intervention group than the control group ($P < 0.001$). This is an improvement in thinking capacity, which is in tandem with the growing global studies. Specifically, one study by Gan et al. (2024) has demonstrated that orthopedic learning with AI implementation has a positive effect on long and short-term scores (20). In addition, the findings of Huang et al. (2024) in the surgical clerkships demonstrated the positive impact of chatbots on the theoretical scores and active participation of learners (21). A meta-analysis by Liu et al. (2025) also found that the RAG technique enhances the

likelihood of correct responding by a medical professional by an average of 1.35 times relative to base models (22); a result that explains the learning leap achieved in this study.

It seems that the concept of the RAG model as an Intelligent Teaching Assistant has taken the learning process that was previously in its passive form and evolved into an active, inquiry-driven process. The same findings of Wu et al. (2024), who reported that even the conceptual test scores can be improved with RAG by up to 15% (23), are also in agreement with the quantitative findings of the current research.

4.3. Mechanism of Action: The Role of RAG Architecture in Managing Cognitive Load

The effectiveness of such an intervention may be explained by the use of the Cognitive Load Theory. Conventional learning settings and scanning of large medical databases tend to increase the extraneous cognitive load. Instead, with the help of accurate and fine information retrieval, the RAG model removes this excessive load, and thus the student's working memory capacity is freed to do some deep processing and create mental schemas.

In addition, research conducted by Wang et al. (2024) based on the RISE framework revealed that RAG can greatly enhance the correctness and completeness of responses (24). In the current study, the limited scope of the language model to the surgical technology reference textbooks guaranteed that the information that was given had a logical organization and scientific consistency. This results in the development of a more effective learning process due to the improvement of comprehensiveness and relevance of responses (25).

4.4. Critical Analysis: Superiority over General Models and the Hallucination Challenge

One of the distinguishing features of this study is its overcoming of the "Hallucination" challenge common in general-purpose language models. In their comparison of the specialized chatbot Anatbuddy and ChatGPT 3.5, Arun et al. (2024) have shown that general models produce fluent responses but are likely to make scientific errors (26). The same statement was made by Miao et al. (2024) in the sphere of nephrology, which stated that general models answer generic and sometimes misleading questions, but reference-related models (RAG) have high adherence to clinical guidelines (27).

Nevertheless, a prudent course of action is justified. Even RAG models, Xian et al. (2024) and Zhang et al. (2025) have cautioned, can be prone to Hallucination in complicated medical conditions up to a rate of 10% (28, 29). Even though this study tried to address this risk by limiting the Knowledge Base, the results of Kim (2025) indicate that the RAG technique may not be enough on its own, and it may be further enhanced with the help of so-called Fine-tuning to achieve better outcomes in closed-ended and multiple-choice questions (30).

4.5. User Satisfaction Analysis Based on the Technology Acceptance Model (TAM)

The mean satisfaction reported (72.48) is high, which shows that the students have adopted the technology successfully. According to the Technology Acceptance Model (TAM), the two elements of the concept of Perceived Usefulness and Perceived Ease of Use are the key drivers of this satisfaction. According to Ouyang et al. (2024) and Huang et al. (2024), it was found that instant access to information and a conversational interaction make the process of learning enjoyable among the students of surgery (20, 21).

The key element here is "Trust." As demonstrated by Kaczmarczyk et al. (2025), RAG systems establish trust in the user through a valid citation (31). In the present study, the capability of the model to refer to surgical references probably minimized anxiety due to the lack of knowledge about the validity of the content, which increased the total satisfaction. However, Khakpaki (2025) also indicates that students with lower digital literacy may have lower levels of satisfaction and learning (32), a point that may explain the difference in satisfaction levels of individual students.

4.6. Limitations of the Study

The results have limitations to the interpretation and generalization of the results, which must be approached carefully:

1. **Limitations in the Study Design and Sampling**: The first limitation of this study was the non-random assignment of groups in terms of university locations, which was done to avoid contamination bias in students. Although such practice might cause some concern about the differences in environmental aspects, both universities are required to adhere to the same National Curriculum approved by the Iranian Ministry of Health and Medical Education in the Orthopedic Surgical Technology course. Additionally, both groups had the same reference textbooks, and preliminary statistical tests showed no significant difference in the knowledge of the baseline (pre-test scores) and the demographics. Also, a relatively small sample size (50 individuals) and convenience sampling should be considered as a warning when making conclusions about the whole population of perioperative nursing students in the country. Even though the effect sizes were large, the sample size was rather small, so the intervention effect could have been overestimated, and the effect should be replicated in larger samples.
2. **Focus Solely on the Cognitive Domain**: The evaluation tool applied to this research was restricted to multiple-choice tests (theoretical knowledge). Direct evaluation was not done on practical skills, real-time clinical reasoning, and performance in the actual operating room setting, which is the essence of this discipline. According to JY Kim et al. (2025), RAG might not be as effective in imparting practical skills as opposed to theoretical knowledge (30).
3. **Intervention Duration and Retention**: The study was carried out in one academic semester, and no long-term effects of the intervention were measured. It is not clear whether the knowledge gained with the help of AI can be held at the required balance in the long term (e.g., 6 months or a year).
4. **Lack of Control for Digital Literacy**: The level of digital competence and prior experience with AI was not measured and controlled as a covariate, a factor that can influence the quality of contact and education.

5. **Novelty Effect**: The initial appeal and novelty of the technology might be the cause of some of the high motivation and satisfaction of the students, and will not be maintained in the long-term use.

5. Conclusion

Finally, the results of this study are preliminary empirical evidence that the incorporation of the Retrieval-Augmented Generation (RAG) architecture in the education of Orthopedic Surgical Technology is an effective method of improving the quality of learning and making students more satisfied with the education. This technology has managed to fill the gap between the static knowledge (reference textbooks) and the dynamic educational requirements and to surpass the constraints of general models like Hallucination and vagueness, and has been able to act as a Trusted Educational Assistant. The findings show that the shift to the new paradigm of education based on AI interactive learning allows not only to enhance the level of theoretical knowledge, but also to make the educational process more interesting, as it helps to reduce cognitive load and provides real-time feedback. Although the current research has certain limitations in the evaluation of practical skills, the study can be regarded as a strong foundation for the intelligent implementation of new tools into the curriculum of the surgical sciences and the need to shift towards Hybrid Learning models.

Abbreviations

AI: Artificial Intelligence; LLM: Large Language Model; RAG: Retrieval-Augmented Generation; VR: Virtual Reality; CVR: Content Validity Ratio; CVI: Content Validity Index; SPSS: Statistical Package for the Social Sciences.

Declarations

Ethics approval and consent to participate

This study was carried out in accordance with the Declaration of Helsinki. It was approved by the Iran University of Medical Sciences Ethics Committee (IR.IUMS.FAM.REC.1404.017). Informed consent was obtained from all the study subjects before the start of the study.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by the Vice Chancellor for Research and Technology, Iran University of Medical Sciences.

Authors' contributions

B.P. designed the study, carried out the study, and wrote the manuscript. Z.B. designed the software and the RAG system. N.A. carried out the statistical analysis. S.H. and F.A. supervised the study, revised the manuscript, and approved the manuscript.

Acknowledgements

The authors would like to thank the Vice-Chancellor of Research and Technology at Iran University of Medical Sciences and the officials at AJA University of Medical Sciences for their cooperation. We also thank the students and experts who participated in this study.

Declaration of Generative AI and AI-assisted technologies in the writing process

As the language of the authors is not English, the use of Generative AI and Large Language Model (LLM) tools was applied when preparing this work only to enhance its readability, language accuracy, and syntax. These technologies could only be used in language refinement and proofreading. The authors revised and edited when necessary and have the full responsibility for the content of the publication.

References

1. Lentz CM, De Wijngaarden L, Willeboordse RAF, Hooft F, van der Laan L. MJ. Dedicated teams to optimize quality and safety of surgery: A systematic review. Int J Qual Health Care. 2022;34(4).
2. Tueni N, Amirouche F. Branding a New Technological Outlook for Future Orthopaedics. Bioengineering. 2025;12(5):494.
3. Sameera V, Bindra A, Rath GP. Human errors and their prevention in healthcare. J Anaesthesiol Clin Pharmacol. 2021;37(3):328–35.
4. Sbitan L, Alzraikat N, Tanous H, Saad A, Odeh M. From one size fits all to a tailored approach: integrating precision medicine into medical education. BMC Med Educ. 2025;25.
5. Hart S, Mind. Brain Educ. 2016;10.
6. Yang R, Fu M, Tantithamthavorn C, Arora C, Vandenhurk L, Chua J. RAGVA: Engineering retrieval augmented generation-based virtual assistants in practice. J Syst Softw. 2025;226:112436.
7. Champagne BJ. Effective teaching and feedback strategies in the or and beyond. Clin Colon Rectal Surg. 2013;26(4):244–9.

8. Bagheri M, Shiryani M, Mehralizade S, Basirinezhad MH, Bagheri H, Abbasi A. The effect of educational intervention using surgery preference cards on anxiety and clinical competence among operating room nursing students in general surgeries: a randomized controlled trial study. BMC Med Educ. 2025;25(1):1338.
9. Tokuno J, Carver TE, Fried GM. Measurement and Management of Cognitive Load in Surgical Education: A Narrative Review. J Surg Educ. 2023;80(2):208–15.
10. Valente DS, Brasil LJ, Spinelli LF, Vilela MAP, Rhoden EL. A narrative review of transforming surgical education with artificial intelligence: opportunities and challenges. AME Surg J. 2025;5.
11. Hla DA, Hindin DI, Generative. AI & machine learning in surgical education. Curr Probl Surg. 2025;63:101701.
12. Sajjadi M, Kara B, Eyupoglu C, Karakuş O. A Survey on Hallucination in Large Language Models: Definitions, Detection, and Mitigation2025.
13. Ramachandran A. Retrieval Augmented Generation (RAG) for Large Language Models Leveraging Enterprise Data (SAP, Salesforce, Workday). 2024.
14. Li Z, Wang Z, Wang W, Hung K, Xie H, Wang FL. Retrieval-augmented generation for educational application: A systematic survey. Computers Education: Artif Intell. 2025:100417.
15. Ntakakis G, Plomariti C, Frantzidis C, Antoniou PE, Bamidis PD, Tsoulfas G. Exploring the use of virtual reality in surgical education. World J Transpl. 2023;13(2):36–43.
16. Ejaz A, Farhan M, Alessi Longa F. AI and Cognitive Load: How Reliance on AI Tools (Chatgpt, Etc.) Affects Critical Thinking. Res J Psychol. 2025;3.
17. Đerić E, Frank D, Milković M. Trust in Generative AI Tools: A Comparative Study of Higher Education Students, Teachers, and Researchers. Information. 2025;16:622.
18. MIR MM, MIR GM, RAINA NT, MIR SM, MIR SM. Application of Artificial Intelligence in Medical Education: Current Scenario and Future Perspectives. J Adv Med Educ Professionalism. 2023;11(3):133–40.
19. Baur D, Ansorg J, Heyde CE, Voelker A. Development and Evaluation of a Retrieval-Augmented Generation Chatbot for Orthopedic and Trauma Surgery Patient Education: Mixed-Methods Study. Jmir ai. 2025;4:e75262.
20. Gan W, Ouyang J, Li H, Xue Z, Zhang Y, Dong Q, et al. Integrating ChatGPT in orthopedic education for medical undergraduates: randomized controlled trial. J Med Internet Res. 2024;26:e57037.
21. Huang Y, Xu B-b, Wang X-y, Luo Y-c, Teng M-m, Weng X. Implementation and evaluation of an optimized surgical clerkship teaching model utilizing ChatGPT. BMC Med Educ. 2024;24(1):1540.
22. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inf Assoc. 2025;32(4):605–15.
23. Wu S, Xiong Y, Cui Y, Wu H, Chen C, Yuan Y et al. Retrieval-augmented generation for natural language processing: A survey. arXiv preprint arXiv:240713193. 2024.

24. Wang D, Liang J, Ye J, Li J, Li J, Zhang Q, et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: comparative study. J Med Internet Res. 2024;26:e58041.
25. Aytar AY, Kaya K, Kılıç K. A synergistic multi-stage RAG architecture for boosting context relevance in data science literature. Nat Lang Process J. 2025:100179.
26. Arun G, Perumal V, Urias FPJB, Ler YE, Tan BWT, Vallabhajosyula R, et al. ChatGPT versus a customized AI chatbot (Anatbuddy) for anatomy education: A comparative pilot study. Anat Sci Educ. 2024;17(7):1396–405.
27. Miao J, Thongprayoon C, Suppadungsuk S, Garcia Valencia OA, Cheungpasitporn W. Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications. Medicina. 2024;60(3):445.
28. Xian X, Wang G, Bi X, Srinivasa J, Kundu A, Fleming C et al. On the Vulnerability of Applying Retrieval-Augmented Generation within Knowledge-Intensive Application Domains2024 September 01, 2024: [arXiv:2409.17275 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2024arXiv240917275X
29. Zhang W, Zhang J. Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review. Mathematics. 2025;13(5):856.
30. Kim S, Medbiolm. Optimizing medical and biological qa with fine-tuned large language models and retrieval-augmented generation. arXiv preprint arXiv:250203004. 2025.
31. Kaczmarek JI, Pokrywka J, Biedalak K, Kurzyp G, Grzybowski Ł. Optimizing Retrieval-Augmented Generation of Medical Content for Spaced Repetition Learning. arXiv preprint arXiv:250301859. 2025.
32. Khakpaki A. Advancements in artificial intelligence transforming medical education: a comprehensive overview. Med Educ Online. 2025;30(1):2542807.

Figures

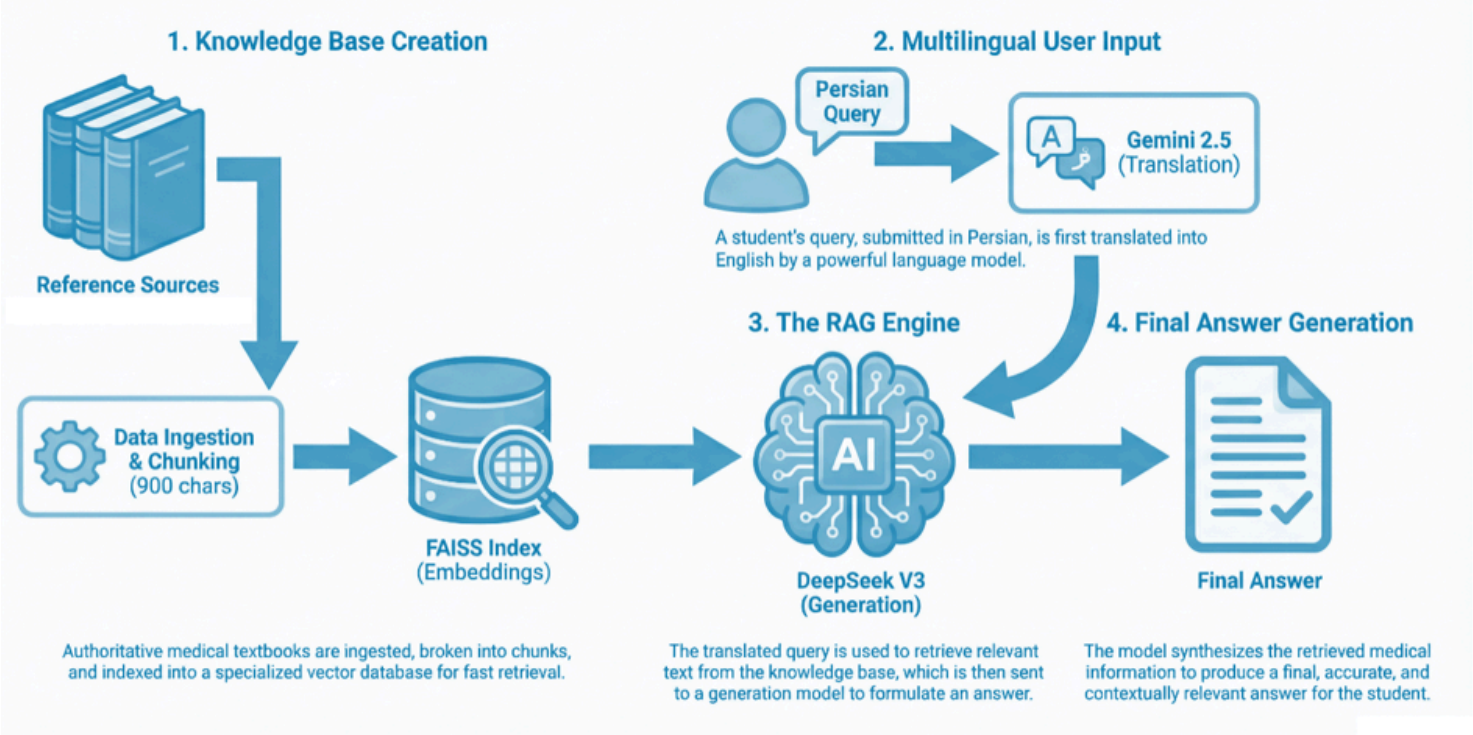


Figure 1

Schematic diagram of the Domain-Specific RAG Architecture used in the study.

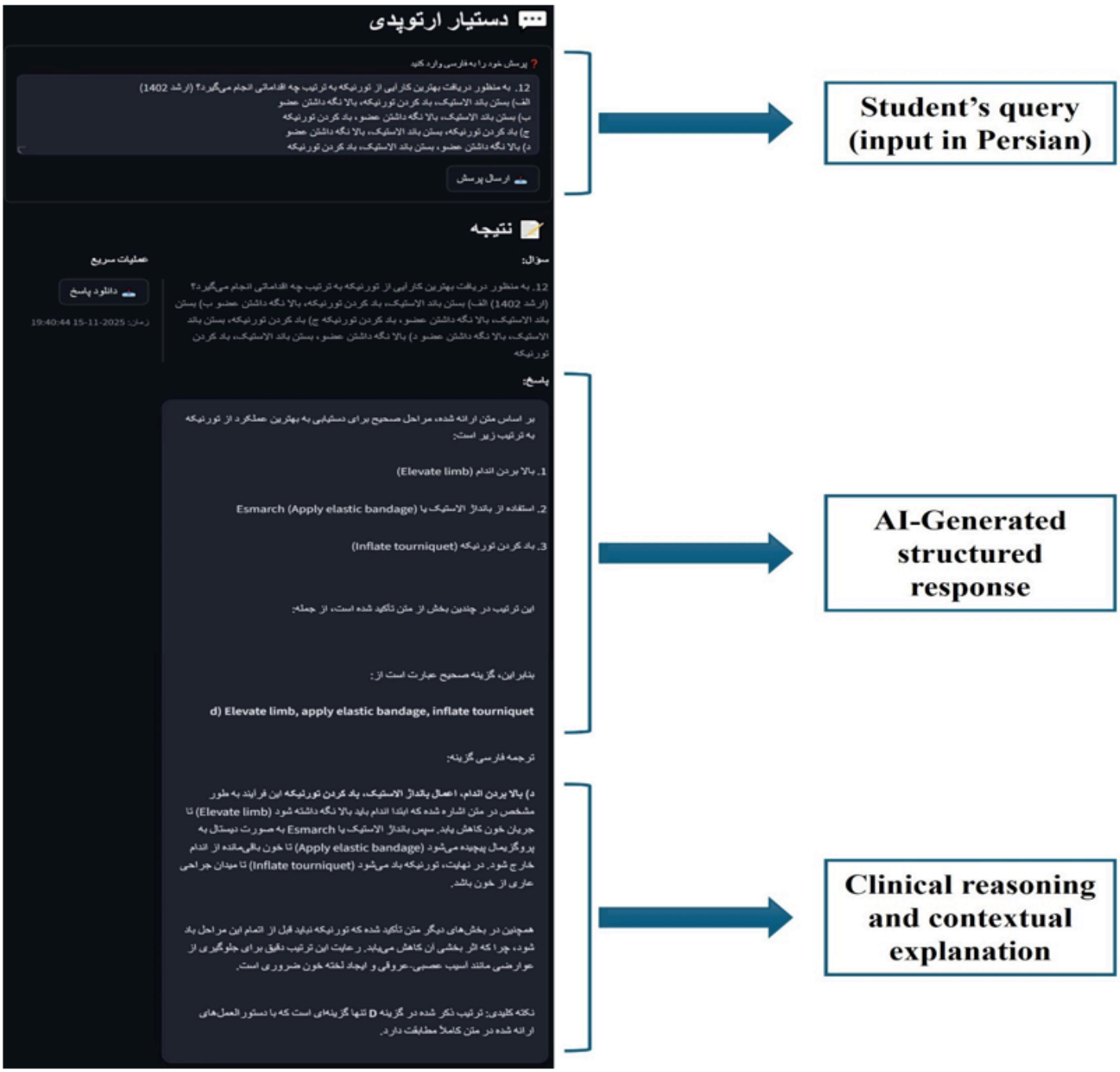


Figure 2

Intuition of the RAG-Based Chatbot developed. The character shows an example of a real-life interaction in the learning platform (built with Streamlit). The process begins with the query entry of the student in Persian (top), to the structured AI-generated response of the system on the basis of reference textbooks (middle), and then a detailed explanation of clinical reasoning and contextual explanation to strengthen learning (bottom).

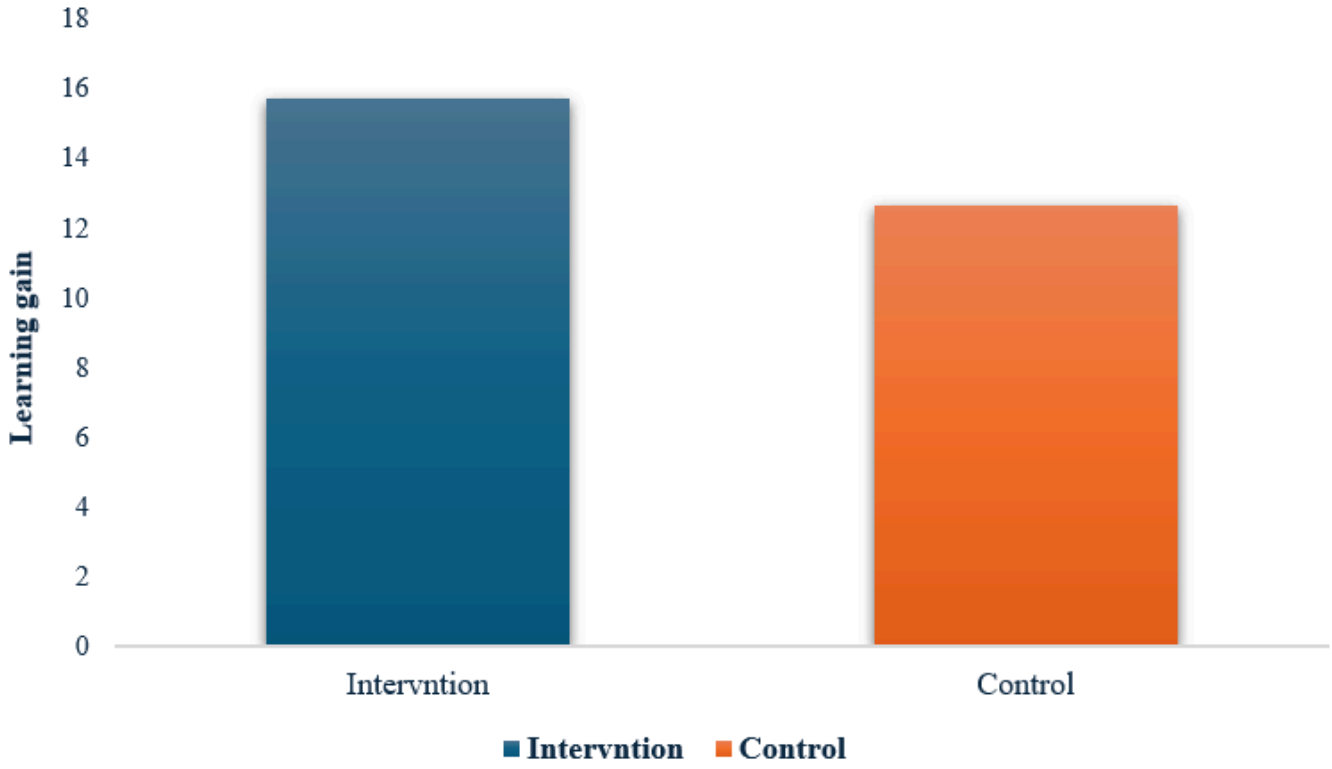


Figure 3

Comparison of mean learning gain scores between the intervention (RAG Model) and control (Conventional) groups (Mann-Whitney U test, $P < 0.001$).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile1.docx