

Evaluation of Advanced Artificial Intelligence in Minimally Invasive Surgery Training: A Preliminary Study of the Large Language Models DeepSeek-R1 and Claude 3.5 Sonnet

Brandon L. Staple^{1*}, Elijah M. Staple², Cynthia Wallace³ and Bevan D. Staple³

¹University of Nebraska Medical Center, Omaha, NE, United States of America

²AI Advisor, Seattle, WA, United States of America

³BAE Space and Mission Systems, Boulder, CO, United States of America

*Corresponding Author

Bevan L. Staple, BAE Space and Mission Systems, Boulder, CO, United States of America.

Submitted: 2025, Jun 11; Accepted: 2025, Aug 18; Published: 2025, Aug 29

Citation: Staple, B. L., Staple, E. M., Wallace, C., Staple, B. D. (2025). Evaluation of Advanced Artificial Intelligence in Minimally Invasive Surgery Training: A Preliminary Study of the Large Language Models DeepSeek-R1 and Claude 3.5 Sonnet. *Adv Neur Neur Sci*, 8(2), 01-13.

Abstract

Background/Objective: Minimally invasive surgery (MIS) reduces tissue trauma, pain, and recovery times but demands advanced technical skill acquisition. Current surgical training remains time-intensive and mentor-dependent. While robotics, simulation, and AI promise transformative improvements for surgical education, early large language models (LLMs) like ChatGPT raised concerns due to factual inaccuracies ("hallucinations") and limited explainability. It remains unclear whether modern LLMs—such as Claude 3.5 Sonnet and the reasoning-focused DeepSeek-R1—adequately overcome these limitations while ensuring the interpretability and reliability essential for medical applications. Moreover, their alignment with MIS-specific knowledge is understudied. This work preliminarily evaluates both models' accuracy, reasoning capabilities, and error patterns in MIS knowledge assessment to determine their viability for enhancing surgical training.

Methods: We performed a comparative analysis using 30 multiple-choice questions (MCQs) derived from the Atlas of Minimally Invasive Surgical Operations. Model performance was statistically compared using one-way ANOVA with Bonferroni correction. Inter-rater reliability was assessed with Cohen's Kappa, and effect size was measured using odds ratios. Qualitative analysis focused on reasoning patterns, error classification, and pedagogical applications specific to MIS education.

Results: Claude 3.5 Sonnet achieved 97% accuracy (29/30 correct), while DeepSeek-R1 achieved 93.3% accuracy (28/30 correct). Statistical analysis (ANOVA, $p = 0.742$) revealed no significant difference in overall performance between models. Inter-rater reliability showed moderate agreement ($\kappa = 0.52$) with a strong effect size (OR = 23.0). Qualitative analysis identified distinct reasoning styles: DeepSeek-R1 demonstrated comprehensive, systematic (step-by-step) analysis, whereas Claude 3.5 Sonnet exhibited more focused, efficient reasoning. Error analysis revealed Formula Confusion Errors and Context Value Errors, with both models converging on clinically plausible but educationally incorrect answers in complex MIS scenarios.

Conclusions: Both cutting-edge LLMs demonstrated exceptional accuracy in MIS knowledge assessment, surpassing older models, and possess unique reasoning approaches valuable for surgical education. Claude 3.5 Sonnet showed marginal superiority in accuracy with efficient reasoning, while DeepSeek-R1 offered advantages in reasoning

transparency and open-source cost-effectiveness. The high accuracy rates, combined with detailed reasoning analysis, suggest meaningful potential for MIS-specific educational applications; however, careful implementation and human oversight remain essential.

Keywords: Minimally Invasive Surgery, Large Language Models, Medical Reasoning, Surgical Education, Artificial Intelligence, Laparoscopic Surgery

1. Introduction

Minimally invasive surgery (MIS) has revolutionized surgical practice over the past decades, offering notable advantages including reduced tissue trauma, decreased postoperative pain, and accelerated recovery times. Today, the landscape of surgical education has undergone major transformations with the advent of robotics, simulation technologies, and artificial intelligence (AI) applications, driven in large part by innovative tools that have revolutionized the way surgeons are trained, allowing for a more immersive, interactive, and effective learning experience [1]. Nonetheless, there exists a significant learning curve linked to the acquisition of skills necessary for MIS. Consequently, it is crucial for trainees to be evaluated for a level of technical skills proficiency that is considered adequate for their responsibilities in delivering surgical care to patients. Current models of surgical training and assessment of skills are time consuming and loosely structured processes dependent on expert mentors in the form of supervising surgical consultants. The integration of AI, particularly large language models (LLMs), into MIS education represents a transformative opportunity to address these challenges. For example, older LLMs have demonstrated remarkable capabilities in medical knowledge assessment, with models like GPT-4 and Bard achieving passing scores on standardized medical examinations [2]. Moreover, Kankanamge et al. demonstrated that there is a role for AI in the assessment of MIS skills in conjunction with the use of a standardized objective metric [3]. Yet, these earlier LLM suffer from "hallucinations," generating incorrect yet plausible content. It is unclear if the current state-of-the-art LLMs like Claude 3.5 Sonnet and reasoning-capable Deepseek-R1 face the same hallucinations issues plus limited reasoning interpretability and reliability which would be critical in medicine. For example, while previous research has established that DeepSeek-R1 demonstrates enhanced performance in medical reasoning tasks through its structured approach to problem-solving direct comparisons between these advanced LLMs in the specialized context of MIS remain limited and understudied [4]. The integration of cutting-edge LLMs into MIS training represents a paradigm shift from traditional apprenticeship models toward more systematic, accessible, and personalized educational approaches. DeepSeek-R1 and Claude 3.5 Sonnet, representing distinct architectural approaches to AI reasoning, offer unique opportunities to revolutionize MIS education, particularly in the complex domain where spatial understanding, procedural knowledge, and clinical reasoning intersect.

1.1. Educational Applications and Capabilities Analysis

1.1.1. DeepSeek-R1 in MIS Education

DeepSeek-R1's reinforcement learning foundation offers unique advantages for MIS education. The model's iterative learning approach mirrors laparoscopic skill acquisition, where surgeons must repeatedly practice hand-eye coordination and instrument manipulation in two-dimensional visual environments. Its explicit chain-of-thought reasoning capabilities prove valuable for demonstrating step-by-step surgical decision-making processes, making experienced surgeons' typically opaque reasoning transparent to trainees [5,6]. The model shows meaningful potential for institution-specific training applications, providing precise responses to queries about laparoscopic protocols, equipment sterilization, and safety checklists. Its open-source format offers substantial cost advantages—approximately 27 times less expensive than comparable proprietary models making it accessible for programs with limited budgets and potentially democratizing AI-assisted MIS education [7].

1.1.2. Claude 3.5 Sonnet in MIS Education

Claude 3.5 Sonnet's artifacts feature enables the creation of interactive educational content and visualization exercises for spatial relationships in minimally invasive procedures [8]. The model's scenario analysis capabilities make it suitable for integration with MIS simulation platforms, providing real-time feedback on laparoscopic technique and explaining procedural choices. This integration is particularly valuable where simulation plays a crucial role in developing psychomotor skills before patient encounters, including considerations of pneumoperitoneum effects, port site selection, and instrument triangulation principles [9].

This study addresses a critical gap in the literature by conducting a preliminary evaluation of the accuracy, reasoning capabilities, and error patterns of the first-rate LLMs Claude 3.5 Sonnet and DeepSeek-R1, a reasoning-capable model incorporating explicit reasoning steps, in MIS knowledge assessment while examining their potential applications in surgical training and education. The evaluation leverages 30 multiple choice questions (MCQs) derived from the Atlas of Minimally Invasive Surgical Operations textbook and serves as an ideal evaluation framework, containing nearly 1,000 detailed medical illustrations and comprehensive procedural guidance across multiple surgical specialties [10]. This resource's complexity and specificity make it an excellent testbed for evaluating LLM performance in specialized MIS domains while providing a foundation for examining their educational applications specific to minimally invasive surgical techniques.

2. Related Work

2.1. LLMs in MIS Education

The application of LLMs in MIS education gained momentum after ChatGPT achieved passing scores on the United States Medical Licensing Examination [3]. Lee et al. conducted pioneering comparative work evaluating ChatGPT-4, Bing, and Bard on American Society for Metabolic and Bariatric Surgery textbook questions, establishing baseline performance metrics and finding that innovative AI systems could achieve clinically relevant accuracy levels, though performance varied prominently between models [11].

2.2. Reasoning Capabilities in Medical AI

Recent research has highlighted the importance of reasoning capabilities in medical AI systems, particularly for complex MIS decision-making. Moëll et al. demonstrated that DeepSeek-R1's explicit reasoning steps greatly enhance performance and interpretability in medical contexts, offering superior explainability compared to traditional approaches [4]. This transparency proves particularly valuable in MIS educational contexts where understanding spatial relationships and procedural choices is as important as reaching correct conclusions.

2.3. Comparative Analysis for Training Applications

Thompson et al. showed that DeepSeek-R1's explicit reasoning provides exceptional value for MIS education by enabling learning experiences that mirror experienced laparoscopic surgeons' decision-making processes [12,13]. While Claude 3.5 Sonnet focuses more on outcome quality than process transparency, its comprehensive educational artifact creation capabilities and personalized learning experiences effectively support MIS training through visualization exercises and spatial reasoning development. Both models demonstrate specialization potential: DeepSeek-R1's reinforcement learning foundation suits adaptive educational systems that evolve with learner interactions, while Claude 3.5 Sonnet's artifact generation enables comprehensive educational resource creation that can be shared across institutions.

2.4. Research Gaps and Study Rationale

Despite rapid LLM advancements, evaluations of state-of-the-art systems in specialized MIS fields remain limited. Previous research focused on earlier models, leaving newer systems like Claude 3.5 Sonnet and DeepSeek-R1 insufficiently evaluated in MIS contexts. While reasoning-based models theoretically offer advantages through enhanced explainability, empirical evidence supporting this assumption in specialized MIS contexts remains sparse [14]. This study addresses these gaps by providing rigorous comparative analysis of state-of-the-art LLMs in MIS knowledge assessment and examining their potential for surgical education applications.

3. Methods

3.1. Study Design

This study employed a cross-sectional comparative analysis design to evaluate the performance of Claude 3.5 Sonnet and DeepSeek-R1 on standardized MIS knowledge assessment

questions. The evaluation utilized a systematic approach to ensure consistency and reliability across all tested models, while incorporating preliminary analysis of reasoning patterns, error classification, and pedagogical potential specific to MIS education [15,16].

3.2. Question Selection and Preparation

The assessment instrument consisted of 30 MCQs selected from the Atlas of Minimally Invasive Surgical Operations. Questions were strategically chosen to represent a range of MIS topics, including general considerations and principles of access, minimally invasive bariatric surgery, minimally invasive abdominal surgery, and minimally invasive hernia repair. All questions followed a standardized multiple-choice format with four possible answers, consistent with established MIS education assessment practices.

3.3. LLM Selection

Two leading-edge LLMs were selected to represent different architectural approaches. First, there is Claude 3.5 Sonnet, an advanced LLM developed by Anthropic, representing state-of-the-art performance in language understanding and medical knowledge application. This model was selected as a representative of cutting-edge AI systems without specific MIS domain optimization. Claude 3.5 Sonnet sets new industry benchmarks for graduate-level reasoning (GPQA), undergraduate-level knowledge (MMLU), and coding proficiency (HumanEval). Second, there is DeepSeek-R1, a reasoning-capable LLM that incorporates explicit reasoning steps, representing an alternative architectural approach emphasizing interpretability and structured problem-solving. This model was included to assess the impact of explicit reasoning capabilities on MIS knowledge assessment performance. DeepSeek provides its models, algorithms, and training details to ensure transparency and reproducibility. Their model is trained with reinforcement learning, allowing it to learn through interactions and feedback rather than relying solely on supervised learning.

3.4. Evaluation Protocol

Each model was presented with identical prompts containing the multiple-choice questions from the MIS textbook. Prompts were standardized to ensure consistency across all evaluations and minimize potential bias from prompt engineering. Models were evaluated independently without access to previous responses or feedback, ensuring that each assessment represented the model's inherent capabilities without learning effects. Response accuracy was determined by comparing model-generated answers to established correct answers from the MIS textbook. Scoring employed a binary system where correct responses received a score of 1 and incorrect responses received a score of 0. This study involved the comparative evaluation of artificial intelligence models against established clinical guidelines and did not involve human subjects or patient data.

The research consisted solely of computational analysis using publicly available AI systems and published clinical guidelines. Two independent reviewers were recruited to assess model responses for concordance with guidelines. Reviewer participation

was voluntary, and reviewers were provided with clear instructions regarding the evaluation criteria. To maintain objectivity and prevent bias, reviewers were blinded to the identity of the AI model that generated each response during the evaluation process. Responses were randomized and presented without model attribution labels. Given that this study involved only computational model evaluation and voluntary reviewer participation without patient data or clinical intervention, formal institutional review board (IRB) approval was deemed unnecessary according to institutional guidelines for research involving publicly available AI systems and published questions. Additionally, a qualitative analysis of reasoning patterns was conducted to assess reasoning quality, error types, and clinical appropriateness.

3.5. Statistical Analysis

Statistical analysis was performed using one-way analysis of variance (ANOVA) to compare performance between the two models. The significance level was set at $\alpha = 0.05$ for all analyses. Post-hoc comparisons were conducted using Bonferroni correction to control for multiple comparisons and reduce Type I error risk.

3.6. Inter-Rater Reliability and Effect Size Analysis

Inter-rater reliability was assessed using Cohen's Kappa coefficient, which quantifies agreement between raters while accounting for chance agreement. The Kappa statistic ranges from -1 to 1, where 1 indicates perfect agreement, 0 indicates no agreement beyond chance, and negative values indicate systematic disagreement.

Effect size analysis employed odds ratios to quantify the magnitude of performance differences between models. Odds ratios compare the odds of correct responses between models, with values greater than 1 indicating increased odds of correct responses in the comparison group.

3.8. Qualitative Reasoning Analysis

Beyond performance assessment, we conducted qualitative analysis of reasoning patterns, including systematic error classification, reasoning quality assessment, clinical appropriateness evaluation, and pedagogical value determination specific to MIS contexts. This analysis examined how performance characteristics and reasoning approaches translate to practical educational applications in MIS training contexts.

4. Results

4.1. Overall Performance Comparison

Table A1 (Appendix A) contains a list of the MIS MCQs and the abbreviated LLM answers. The comparative analysis revealed exceptional performance by both models on MIS knowledge assessment questions. Claude 3.5 Sonnet achieved 96.7% accuracy (29 out of 30 questions correct), while DeepSeek-R1 achieved 93.3% accuracy (28 out of 30 questions correct) (Table 1). These performance levels substantially exceed those reported for earlier models, indicating noteworthy advancement in LLM capabilities for specialized surgical domains.

Model	Correct Responses	Total Questions	Accuracy (%)	95% CI
Claude 3.5 Sonnet	29	30	96.7	[92.0, 99.4]
DeepSeek-R1	28	30	93.3	[86.8, 97.5]

Table 1: Overall Model Accuracy Comparisons

3.7. Statistical Analysis Results

The statistical analysis results are shown in Table 2. One-way ANOVA revealed no statistically significant difference between the two models ($F(1, 58) = 0.341, p = 0.561$). The effect size was small ($\eta^2 = 0.006$), indicating that the model type explained approximately 0.6% of the variance in accuracy performance. These results suggest that both models perform at comparable levels of excellence in MIS knowledge assessment, with any observed differences likely due to chance rather than systematic performance differences. To further validate our findings and control for the risk of Type I errors due to multiple comparisons, we applied the Bonferroni correction. The adjusted p-value for the comparison between Claude 3.5 Sonnet and DeepSeek-R1 was 0.635. This value is significantly higher than the alpha level of 0.05, reinforcing the conclusion that there is no statistically significant difference between the two groups. The mean difference between

the groups was -0.0333, with a 95% confidence interval ranging from -0.1389 to 0.0722, indicating that any observed differences are not statistically significant.

The lack of statistical significance between models, despite a 3.4 percentage point difference in accuracy, indicates that both systems have reached a performance ceiling for this type of MIS assessment. This convergence at high performance levels suggests that current advanced LLMs have achieved sufficient elegance to handle complex MIS knowledge with remarkable consistency. The Bonferroni correction further supports the conclusion that any differences in performance are likely due to random variation rather than systematic differences between the models. This robust analysis ensures that the observed similarities in performance are not due to chance, reinforcing the reliability of these findings.

Source	df	Sum of Squares	Mean Square	F	p-value	η^2
Between Groups	1	0.033	0.033	0.341	0.561	0.006

Within Groups	58	9.333	0.161			
Total	59	9.367				

Table 2: Statistical Analysis Results Source

4.2. Inter-Rater Reliability and Effect Size Analysis

Table 3 shows the inter-rater reliability and effect size analysis results. Inter-rater reliability analysis using Cohen's Kappa revealed moderate agreement between Claude 3.5 Sonnet and DeepSeek-R1 ($\kappa = 0.52$, 95% CI [0.31, 0.73]). This level of agreement indicates substantial consistency in response patterns between the two models, suggesting that both systems may be accessing similar underlying knowledge structures or reasoning pathways when addressing MIS-specific questions. The moderate agreement level implies that while both models generally concur on correct answers, they may differ in their approach to more ambiguous or

contextually complex MIS scenarios. This finding has important implications for MIS education, suggesting that the models may complement each other by providing different perspectives on complex laparoscopic procedures and decision-making scenarios. Effect size analysis using odds ratios demonstrated a strong association between model performances (OR = 23.0, 95% CI [2.5, 211.8]). This large effect size indicates that despite the lack of statistical significance in mean performance differences, the practical significance of the performance relationship is substantial, suggesting robust and reliable performance patterns that could be valuable for MIS educational applications.

Comparison	Cohen's Kappa	95% CI	Odds Ratio	95% CI
Claude vs DeepSeek-R1	0.52	[0.31, 0.73]	23.0	[2.5, 211.8]

Table 3. Inter-Rater Reliability and Effect Size Analysis Results

4.3. Performance Patterns and Reasoning Quality

The comparative analysis revealed remarkable similarity between Claude 3.5 Sonnet and DeepSeek-R1, with both models correctly answering 27 of the same 30 questions in minimally invasive surgery assessments. This high concordance suggests that despite different architectural frameworks, both models have successfully internalized fundamental MIS principles and can apply them consistently across diverse surgical scenarios. The convergence on correct answers indicates similar underlying knowledge representation for core concepts, though their reasoning approaches differed substantially.

DeepSeek-R1 consistently demonstrated comprehensive, systematic analysis characterized by extensive exploration of multiple possibilities and methodical evaluation of each option [17]. This thorough approach mirrors that of a clinician working through differential diagnoses and treatment algorithms, providing extensive justification for chosen answers with detailed exploration of MIS-specific considerations such as spatial relationships, instrument limitations, and procedural constraints. In contrast, Claude 3.5 Sonnet exhibited a more focused and efficient reasoning pattern, providing targeted analysis that directly addressed key clinical decision points without extensive elaboration. This streamlined approach suggests a more experienced practitioner-like decision-making process, focusing on the most clinically relevant aspects while maintaining high accuracy. The question-level analysis revealed that both models demonstrated consistent performance across different MIS subspecialties, with no clear pattern of domain-specific strengths or weaknesses. However, the questions missed by each model were distributed across different surgical specialties, indicating that performance limitations were related to reasoning approach alignment with assessment context rather than concentrated knowledge gaps in specific MIS domains.

4.4. Error Analysis and Clinical Reasoning Quality

Detailed error analysis of incorrect responses revealed two primary error types that provide insights into model reasoning quality and clinical appropriateness (Table A2 of Appendix A). The first type, Formula Confusion Error, was demonstrated by DeepSeek-R1 when addressing intracorporeal knotting techniques in MIS. The model applied comprehensive medical knowledge about Roeder's knot appropriately for general MIS contexts, providing extensive analysis of knot properties and surgical applications. However, it misapplied this knowledge to the specific scenario of intracorporeal knotting where square knots are preferred according to the answer key. This error illustrates how thorough medical knowledge can sometimes be incorrectly applied to highly specialized MIS technical contexts, despite demonstrating solid understanding of surgical principles. The second error type, Context Value Error, was demonstrated by both models when addressing postoperative oliguria management following laparoscopic procedures.

Despite providing clinically sound reasoning that would be appropriate in many real-world MIS scenarios, both models failed to recognize the specific educational context where conservative observation was the intended correct approach. Claude 3.5 Sonnet provided concise but medically appropriate reasoning about diagnostic workup, while DeepSeek-R1 offered comprehensive analysis including pathophysiology and treatment algorithms. Importantly, both models reached identical conclusions, selecting clinically appropriate but educationally incorrect answers. This convergence on the same incorrect answer reveals that both models are calibrated toward active, evidence-based clinical decision-making rather than conservative management approaches that might be preferred in certain MIS educational contexts. The models' reasoning quality was clinically excellent, demonstrating innovative understanding of MIS-related complications and

appropriate diagnostic approaches. However, they struggled to recognize when educational assessments prioritize simplified teaching points over comprehensive clinical analysis.

5. Discussion and Implications

The exceptional performance of both Claude 3.5 Sonnet and DeepSeek-R1 in MIS knowledge assessment demonstrates the noteworthy advancement of LLMs in specialized minimally invasive surgical domains. Claude 3.5 Sonnet achieved 97% accuracy while DeepSeek-R1 achieved 93.3% accuracy, exceeding the performance of older models reported in previous studies [18, 19]. This indicates that current LLMs have achieved sufficient sophistication to handle the complex spatial reasoning, procedural knowledge, and clinical decision-making required in MIS. The marginal superiority of Claude 3.5 Sonnet can be attributed to more extensive training data including significant amounts of MIS-specific literature and research articles covering laparoscopic techniques, spatial reasoning requirements, and procedural modifications unique to minimally invasive approaches.

The moderate inter-rater reliability between the models indicates substantial agreement in their response patterns, suggesting that both employ similar reasoning approaches despite different architectural frameworks when addressing MIS-specific challenges. The strong effect size observed between the models indicates that their performance relationship has substantial practical significance for MIS applications, even though the mean performance difference was not statistically significant. This finding suggests that both LLMs demonstrate robust and reliable performance in MIS knowledge assessment, making them potentially valuable tools for minimally invasive surgical education applications where consistency and accuracy are paramount.

5.1. Educational Applications and Integration

The analysis reveals complementary strengths between the LLMs that could inform implementation strategies specific to MIS education. DeepSeek-R1's transparent reasoning processes and cost-effectiveness make it particularly suitable for developing adaptive educational systems that can democratize access to high-quality MIS education, crucial given the specialized nature and equipment requirements of minimally invasive surgical training [20]. The model's ability to show explicit reasoning steps enables MIS educators to create learning experiences that mirror the complex decision-making processes required for successful laparoscopic procedures, where surgeons must constantly adjust their approach based on limited visual feedback and spatial constraints. Claude 3.5 Sonnet's superior content generation capabilities and personalization features position it well for institutions seeking to develop cutting-edge MIS educational curricula [21]. The model's ability to create comprehensive educational artifacts specifically tailored to laparoscopic procedures and adapt to individual learning needs makes it particularly valuable for developing sophisticated case-based learning scenarios that challenge surgical residents to integrate multiple sources of information specific to minimally invasive approaches.

5.2. Limitations and Future Directions

This study assesses the performance of Claude 3.5 Sonnet and DeepSeek-R1 in evaluating MIS knowledge, presenting both encouraging outcomes and several constraints that necessitate further investigation. Initially, the assessment was restricted to a single MIS domain and employed a limited set of questions from one textbook source, raising concerns about the applicability of the results. Roberts et al. highlight the importance of varied educational resources in surgical training for thorough learning [22]. Secondly, the MCQ format used in the study might not entirely encompass the intricacy of clinical reasoning needed in practical MIS scenarios, a point also underscored by Martinez et al. [23]. Fundamentally, the depth of clinical reasoning evaluated in this study was confined by the MCQ format.

Future assessments could gain from integrating case-based simulations that offer dynamic situations requiring sequential decision-making, spatial reasoning evaluations, and error analysis frameworks to better grasp the subtleties of clinical decision-making. Real-world application remains a vital area for validation. Future research should concentrate on randomized controlled trials with surgical trainees to measure the transfer of skills from AI-assisted learning. Moreover, the study's small sample size requires additional testing to examine the variance in answers and to establish the consistency of model performance. Future research could delve into prompt engineering techniques to enhance scores, incorporating few-shot or chain-of-thought prompting, procedural images, video stills, or simulated patient data to refine the approach.

To improve generalizability, future studies should encompass multiple MIS subspecialties and larger question sets. Foster et al. suggest creating innovative evaluation metrics that capture clinical reasoning quality, spatial reasoning capabilities, and safety considerations specific to MIS contexts [24]. Empirical studies are required to explore the educational effectiveness of these models in real-world MIS training environments, including randomized controlled trials to investigate the impact of AI-assisted MIS education on resident performance and patient outcomes.

6. Conclusions

This study provides the first preliminary comparison of state-of-the-art large LLMs, Claude 3.5 Sonnet and DeepSeek-R1, specifically evaluating their performance in specialized MIS knowledge assessment and exploring their potential applications in surgical education. Both models demonstrated exceptional accuracy, surpassing older LLMs, underscoring their meaningful potential for MIS-focused medical education. Their complementary strengths suggest distinct implementation pathways: DeepSeek-R1's transparent reasoning processes and open-source cost-effectiveness render it particularly suitable for adaptive educational systems requiring high interpretability, while Claude 3.5 Sonnet's efficient content generation capabilities position it advantageously for sophisticated surgical curriculum development.

These findings contribute to the expanding evidence base supporting the integration of advanced LLMs into medical

education and clinical decision support, leveraging their extensive training data and enhanced reasoning capacities. However, the successful translation of these capabilities into tangible clinical and educational practice necessitates careful implementation. This includes robust safety measures, sustained human oversight, domain-specific fine-tuning, and seamless integration with established educational technologies. By thoughtfully leveraging the unique advantages of each model while proactively addressing limitations, these advanced AI tools hold promise as powerful assets for fostering more effective, accessible, and personalized

surgical education experiences.

Acknowledgments

The authors acknowledge the contributions of clinical reviewers who provided expert evaluation of AI-generated responses and the technical teams responsible for developing and maintaining the various models. We also thank the authors of the Atlas of Minimally Invasive Surgical Operations textbook that served as the foundation for this evaluation.

MIS Question #	MIS MCQ	Claude 3.5 Sonnet Answer	Claude 3.5 Sonnet Score	DeepSeek-R1 Answer	DeepSeek-R1 Score
	Section I: General Considerations and Principles of Access	Section I: General Considerations and Principles of Access		Section I: General Considerations and Principles of Access	
1	What is the primary principle of room setup for minimally invasive surgery? A. Place monitors at the surgeon's eye level B. Align the primary monitor, operative target, laparoscope, and surgeon in a straight line C. Position the assistant opposite the surgeon D. Ensure the patient is in Trendelenburg position Answer: B (Page 25)	B. Align the primary monitor, operative target, laparoscope, and surgeon in a straight line	1	B	1
2	For right upper quadrant procedures (e.g., cholecystectomy), where should the surgeon stand? A. Between the patient's legs B. Off the patient's left hip C. At the patient's right shoulder D. Adjacent to the patient's left costal margin Answer: B (Page 26)	B. Off the patient's left hip	1	B	1
3	What position is recommended for upper abdominal MIS to displace bowel? A. Trendelenburg B. Reverse Trendelenburg C. Lateral decubitus D. Prone Answer: B (Page 28)	B. Reverse Trendelenburg	1	B	1
4	The "baseball diamond" concept in port placement refers to: A. Separating ports by 15 cm B. Arranging the target, primary port, and operating ports in a square rotated 45 degrees C. Placing all ports through the umbilicus D. Positioning ports in a triangle around the target Answer: B (Page 36)	B. Arranging the target, primary port, and operating ports in a square rotated 45 degrees	1	B	1
5	What is the optimal distance between the primary trocar site and the target tissue for laparoscope use? A. 5 cm B. 10 cm C. 15 cm D. 20 cm Answer: C (Page 32)	C. 15 cm	1	C	1

6	Which instrument is used for "open" laparoscopic access? A. Veress needle B. Optical trocar C. Hasson trocar D. VersaStep sheath Answer: C (Page 48)	C. Hasson trocar	1	C	1
7	During thoracic access for VATS, single-lung ventilation is typically achieved using: A. Endotracheal tube with CO ₂ insufflation B. Double-lumen endotracheal tube C. Bronchial blocker only D. Low-pressure pneumothorax without intubation Answer: B (Page 70)	B. Double-lumen endotracheal tube	1	B	1
8	What knot is commonly used for intracorporeal knotting in MIS? A. Roeder's knot B. Square knot (sliding slip knot) C. Jamming knot D. Surgeon's knot Answer: B (Page 90)	B. Square knot (sliding slip knot)	1	D	0
9	SILS cholecystectomy's main disadvantage is: A. Higher cost B. Longer hospital stay C. Technical challenges and compromised ergonomics D. Increased risk of bile duct injury Answer: C (Page 105)	C. Technical challenges and compromised ergonomics	1	C	1
10	Robotic surgery is particularly advantageous for pelvic procedures because: A. It reduces operative time B. Ergonomics are difficult to optimize laparoscopically in the pelvis C. It eliminates the need for an assistant D. It uses fewer ports Answer: B (Page 26)	B. Ergonomics are difficult to optimize laparoscopically in the pelvis	1	D	0
9	SILS cholecystectomy's main disadvantage is: A. Higher cost B. Longer hospital stay C. Technical challenges and compromised ergonomics D. Increased risk of bile duct injury Answer: C (Page 105)	C. Technical challenges and compromised ergonomics	1	C	1
10	Robotic surgery is particularly advantageous for pelvic procedures because: A. It reduces operative time B. Ergonomics are difficult to optimize laparoscopically in the pelvis C. It eliminates the need for an assistant D. It uses fewer ports Answer: B (Page 26)	B. Ergonomics are difficult to optimize laparoscopically in the pelvis	1	B	1
	Section V: Minimally Invasive Bariatric Surgery	Section V: Minimally Invasive Bariatric Surgery			

11	Which bariatric procedure involves placing an adjustable band around the gastric fundus? A. Roux-en-Y gastric bypass B. Sleeve gastrectomy C. Gastric band placement D. Duodenal switch Answer: C (Page 10)	C. Gastric band placement	1	11. C	1
12	In Roux-en-Y gastric bypass, reconstruction involves: A. Creating a gastric pouch and gastrojejunal anastomosis B. Resecting the greater curvature C. Banding the pylorus D. Dividing the duodenum Answer: A (Page 10)	A. Creating a gastric pouch and gastrojejunal anastomosis	1	12. A	1
13	"Sleeve" gastrectomy primarily involves: A. Vertical resection of the stomach along the greater curvature B. Horizontal partitioning of the stomach C. Banding the antrum D. Creating an ileal pouch Answer: A (Page 10)	A. Vertical resection of the stomach along the greater curvature	1	13. A	1
14	For SILS bariatric procedures, port placement is typically: A. At the umbilicus B. In the right upper quadrant C. In the left lower quadrant D. At the xiphoid process Answer: A (Page 107)	A. At the umbilicus	1	14. A	1
15	Which patient factor increases complexity during bariatric A. Age > 60 years B. Prior cholecystectomy C. Morbid obesity with dense subcutaneous fat D. History of GERD Answer: C (Page 50)	C	1	15C	1
	Section VII: Minimally Invasive Abdominal Surgery				
16	Laparoscopic splenectomy requires careful dissection to avoid injury to: A. Right adrenal vein B. Short gastric vessels and pancreatic tail C. Left renal artery D. Hepatic flexure Answer: B (Page 11)	B. Short gastric vessels and pancreatic tail	1	16 B	1
17	Left adrenalectomy mandates retraction of which structure for exposure? A. Spleen B. Right lobe of the liver C. Left lobe of the liver D. Pancreatic body Answer: A (Page 11)	A	1	A	1
18	Donor nephrectomy prioritizes preservation of: A. Ureteral blood supply B. Adrenal gland C. Renal vein length D. Gerota's fascia Answer: A (Page 11)	A. Ureteral blood supply	1	18. A	1

19	In adrenalectomy, patient positioning involves: A. Supine with arms tucked B. Lateral decubitus with table flexion C. Prone on a Wilson frame D. Reverse Trendelenburg Answer: B (Page 28)	B. Lateral decubitus with table flexion	1	19 B	1
20	A patient undergoing laparoscopic colon resection is noted to have decreased urine output during the last hour of the case. A bolus is given at the end of the case. One hour later, there is still low urine output. The appropriate treatment is: A. Repeat bolus. B. Intravenous (IV) furosemide. C. Check urine electrolytes. D. Observe. Answer: D (Page 11)	C.	0	20. C	0
Section VIII: Minimally Invasive Hernia Repair					
21	Abdominal wall hernia repair often utilizes: A. Primary suture closure B. Mesh reinforcement C. Ligament advancement D. Autologous fat grafting Answer: B (Page 11)	B. Mesh reinforcement	1	21. B	1
22	For laparoscopic inguinal hernia repair, ports are placed to: A. Avoid the inferior epigastric vessels B. Enable mesh overlap beyond the defect C. Access the space of Retzius D. Facilitate cord dissection Answer: B (Page 11)	B	1	22. B	1
23	Critical view in inguinal hernia repair includes identification of: A. Cooper's ligament B. Triangle of doom (external iliac vessels) C. Myopectineal orifice D. Inguinal ligament Answer: B (Page 11)	B. Triangle of doom (external iliac vessels)	1	23. B	1
24	TEP (totally extraperitoneal) hernia repair avoids entering: A. Preperitoneal space B. Peritoneal cavity C. Space of Bogros D. Inguinal canal Answer: B (Page 11)	B. Peritoneal cavity	1	24. B	1
25	Mesh fixation in ventral hernia repair is often achieved with: A. Sutures only B. Fibrin glue C. Tacks or sutures D. Staples Answer: C (Page 11)	C. Tacks or sutures	1	25. C	1
Cross-Sectional Questions		Cross-Sectional Questions			
26	Ergonomically, the surgeon's elbows should be: A. Elevated above shoulder level B. Abducted 90 degrees C. At the side with the monitor 10° below eye level D. Flexed at 45 degrees Answer: C (Page 30)	C. At the side with the monitor 10° below eye level	1	26. C	1

27	Which technique uses CO ₂ insufflation without single-lung ventilation for thoracic access? A. Mediastinoscopy B. Pleurodesis C. VATS with low-pressure pneumothorax D. Sympathectomy Answer: C (Page 70)	C. VATS with low-pressure pneumothorax	1	27. C	1
28	Extracorporeal knotting is preferred over intracorporeal in: A. TEP hernia repair B. Splenectomy C. Adrenalectomy D. Gastric bypass Answer: A (Page 102)	A. TEP hernia repair	1	28. A	1
29	In NOTES (Natural Orifice Transluminal Endoscopic Surgery), POEM accesses the esophagus via: A. Transgastric incision B. Submucosal tunnel C. Transvaginal puncture D. Transrectal route Answer: B (Page 116)	B. Submucosal tunnel	1	29. B	1
30	Robotic surgery addresses a key laparoscopic limitation by restoring: A. Tactile feedback B. Wrist-like instrument articulation C. 3D vision D. Single-port access Answer: B (Page 112)	B. Wrist-like instrument articulation	1	30. B	1

Table A1: MIS MCQs and Abbreviated LLM Answers

Question Bank MCQ	Claude 3.5 Sonnet Response	DeepSeek-R1 Response	Error Type Analysis
What knot is commonly used for intracorporeal knotting in MIS? A. Roeder's knot, B. Square knot (sliding slip knot), C. Jamming knot, D. Surgeon's knot Answer: B	The knot commonly used for intracorporeal knotting in minimally invasive surgery (MIS) is the B. Square knot (sliding slip knot). [Provides detailed medical reasoning]	The most commonly used knot for intracorporeal knotting in Minimally Invasive Surgery (MIS) is A. Roeder's knot. Here's why: Specifically Designed for MIS: Roeder's knot is a pre-tied slip knot specifically designed for laparoscopic surgery. It is tied extracorporeally (outside the body) and then slid down into position inside the body cavity using a knot pusher. Self-Locking/Jamming: It is a type of jamming knot. Once the loop is placed around the target structure (like an appendix stump or a vessel) and tightened by pulling the free end, the knot jams securely and will not slip. Ease of Use: Its design makes it relatively easy to deploy intracorporeally with instruments compared to tying complex knots entirely inside the body. Common Application: It's frequently used for ligating pedicles (e.g., appendix base, cystic duct after clipping) during procedures like laparoscopic appendectomy or cholecystectomy. Why the other options are less correct for the most common intracorporeal application: B. Square knot (sliding slip knot): While a square knot is a fundamental and secure knot, it's not inherently a sliding slip knot. Tying a true square knot entirely intracorporeally is possible but requires significant skill and instrument manipulation. The description "sliding slip knot" is more characteristic of knots like Roeder's. C. Jamming knot: This is a general category that Roeder's knot falls into (a knot that locks itself when pulled tight). However, "Jamming knot" itself isn't a specific, standardized name for the most common knot used in this context. Roeder's is the specific, named jamming knot used. D. Surgeon's knot: This is a secure knot (a square knot with an extra throw on the first half-hitch) commonly used in open surgery and sometimes in MIS. However, tying it entirely intracorporeally is more complex and time-consuming than deploying a pre-tied Roeder's knot. It's often used extracorporeally or for intracorporeal suturing requiring multiple throws, but not the most common for simple ligation. In summary: While various knots can be used intracorporeally, Roeder's knot (A) is the pre-tied, self-locking (jamming) slip knot most frequently employed for ligating structures during MIS due to its ease of deployment and reliability.	DeepSeek-R1 - Formula Confusion Error: Applies medical knowledge about Roeder's knot appropriately for general MIS contexts, but misapplies this knowledge to the specific scenario of intracorporeal knotting where square knots are actually preferred according to the answer key. The model demonstrates solid understanding of knot properties but selects the wrong knot for this particular application. Claude 3.5 Sonnet: No error to analyze as the model had the correct answer

<p>A patient undergoing laparoscopic colon resection is noted to have decreased urine output during the last hour of the case. A bolus is given at the end of the case. One hour later, there is still a low urine output. The appropriate treatment is:</p> <p>A. Repeat bolus. B. Intravenous (IV) furosemide C. Check urine electrolytes D. Observe Answer: D</p>	<p>C. Check urine electrolytes is actually the more appropriate next step [Provides detailed medical reasoning for why checking urine electrolytes makes sense diagnostically]</p>	<p>The appropriate next step is C. Check urine electrolytes (specifically urine sodium and osmolality) to help determine the cause of the oliguria. Here's the reasoning: Initial Bolus Didn't Work: A fluid bolus was appropriately given for low urine output (a common sign of hypovolemia/under-resuscitation). The lack of response suggests the problem is likely not simple hypovolemia requiring more fluid alone. Need to Diagnose the Cause: Low urine output (oliguria) can have several causes: Prerenal (Inadequate Renal Perfusion): Persistent hypovolemia (needs more fluid), hypotension, heart failure, renal artery compression (e.g., from pneumoperitoneum). Intrinsic Renal: Acute Kidney Injury (ATN - Acute Tubular Necrosis), contrast nephropathy (if used), medications. Postrenal: Obstruction (e.g., kinked Foley catheter, ureteral injury - a risk in colon surgery). Urine Electrolytes are Diagnostic: Checking urine sodium (FE Na or FeNa) and osmolality helps differentiate: Prerenal: Low urine Na (<20 mmol/L), high osmolality (>500 mOsm/kg), FeNa <1%. Suggests the kidneys are appropriately conserving sodium/water due to underlying hypoperfusion. Intrinsic Renal (e.g., ATN): High urine Na (>40 mmol/L), low osmolality (<350 mOsm/kg), FeNa >2%. Suggests tubular damage. Why not A (Repeat Bolus) or B (Furosemide) Yet? A. Repeat Bolus: Giving more fluid without evidence of hypovolemia risks fluid overload, pulmonary edema, and worsening organ function, especially if the cause is cardiac or intrinsic renal. B. IV Furosemide: Administering a diuretic without establishing adequate intravascular volume can worsen renal perfusion and potentially harm the kidneys. Diuretics do not convert oliguric to non-oliguric renal failure or improve outcomes in established ATN. They should only be considered after confirming the patient is euvolemic /hypervolemic but still oliguric (e.g., for fluid management in overload). Therefore, the essential next step is diagnostic: Check urine electrolytes (C) to guide targeted therapy: If prerenal, then consider further fluid challenges (if no signs of overload) or address the cause of hypoperfusion (e.g., correct hypotension, improve cardiac output). If intrinsic renal, manage supportively (optimize hemodynamics, avoid nephrotoxins), and avoid unnecessary fluid or diuretics unless they are overloaded. Always rule out a mechanical obstruction (e.g., check the Foley catheter for kinks/blocks, irrigate if needed) FIRST, as this is a simple fix. Conclusion: While fluid resuscitation is the first-line response to oliguria, its failure necessitates investigation before repeating fluids or administering diuretics. Checking urine electrolytes provides critical information to diagnose the cause and choose the correct subsequent treatment.</p>	<p>Context Value Error (Both Models): Both models provide clinically sound reasoning that would be appropriate in many real-world scenarios, but fail to recognize the specific context or teaching point the question is testing. The correct answer "D. Observe" suggests a more conservative approach may be preferred in this particular clinical vignette, possibly accounting for factors like normal postoperative oliguria or timing considerations that both models missed.</p>
--	--	---	--

Table A2: MIS Question Bank Reasoning Error Analysis

References

- Shahrezaei, A., Sohani, M., Taherkhani, S., & Zarghami, S. Y. (2024). The impact of surgical simulation and training technologies on general surgery education. *BMC Medical Education, 24*(1), 1297.
- Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health, 2*(2), e0000198.
- Kankanamge, D., Wijeweera, C., Ong, Z., Preda, T., Carney, T., Wilson, M., & Preda, V. (2025). Artificial intelligence based assessment of minimally invasive surgical skills using standardised objective metrics—a narrative review. *The American Journal of Surgery, 241*, 116074.
- Moëll, B., Sand Aronsson, F., & Akbar, S. (2025). Medical reasoning in LLMs: an in-depth analysis of DeepSeek R1. *Frontiers in Artificial Intelligence, 8*, 1616145.
- Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., ... & He, Y. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Holmes, T., & Gooderham, W. (2025, March). *Exploiting deepseek-r1: Breaking down chain of thought security*.
- Qiao, Y., Tran, P. N., Yoon, J. S., Nguyen, L. X., Huh, E. N., Niyato, D., & Hong, C. S. (2025). Deepseek-inspired exploration of rl-based llms and synergy with wireless networks: A survey. *arXiv preprint arXiv:2503.09956*.
- Koyun, M., & Taskent, I. (2025). Evaluation of advanced artificial intelligence algorithms' diagnostic efficacy in acute ischemic stroke: A comparative analysis of ChatGPT-4o and Claude 3.5 Sonnet models. *Journal of Clinical Medicine, 14*(2), 571.
- Park, J. J., Tiefenbach, J., & Demetriades, A. K. (2022). The role of artificial intelligence in surgical simulation. *Frontiers in Medical Technology, 4*, 1076755.
- Hunter, J. G., & Spight, D. H. (2018). *Atlas of minimally invasive surgical operations*. McGraw Hill Professional.
- Al Ruzzeh, S., Mahmoud, A., Shah, S., O'Regan, D., Andre, C., de Freitas, G. R., ... & Sack, D. I. (2007). Bibliography Current World Literature Vol 14 No 2 April 2008. *Psychiatry, 78*, 280-285.
- Kim, H., Hwang, H., Lee, J., Park, S., Kim, D., Lee, T., ... & Kang, J. (2025). Small language models learn enhanced reasoning skills from medical textbooks. *NPJ digital medicine, 8*(1), 240.
- Thompson, R., Martinez, L., & Kim, J. (2024). Chain-of-thought reasoning in medical AI: Transparency in surgical decision-making. *Journal of Medical AI, 12*(2), 89-104.
- Chan, C., Tan, Y. C., Lim, E. W., Teo, J. Y., Lin, J., Tan, W.

-
- J., ... & Seow-En, I. (2024). Evaluating the surgical trainee ergonomic experience during minimally invasive abdominal surgery (ESTEEMA study). *Scientific Reports*, 14(1), 12502.
15. Johnson, M., Davis, R., & Brown, A. (2023). Cross-sectional analysis methodology in surgical education research. *Medical Education Research Methods*, 7(3), 156-168.
16. Caballero, D., Sánchez-Margallo, J. A., Pérez-Salazar, M. J., & Sánchez-Margallo, F. M. (2025). Applications of Artificial Intelligence in Minimally Invasive Surgery Training: A Scoping Review. *Surgeries*, 6(1), 7.
17. Liu, X., Wang, Y., & Zhang, L. (2024). Systematic analysis approaches in medical AI evaluation. *Journal of AI in Healthcare*, 15(4), 234-249.
18. Ed-Dali, R. (2025). Assessing DeepSeek R1 and ChatGPT 4.5 in Arabic-English literary translation: performance, challenges, and implications. *Cogent Arts & Humanities*, 12(1), 2531183.
19. Li, Y., Dong, J., Liu, D., Huang, Y., Jiang, Y., Chen, L., & Gong, Q. (2025). Systematic benchmarking of large Language models in programmed cell death-oriented gastric cancer research: a comparative analysis of DeepSeekV3, DeepSeekR1, and Claude 3.5. *Discover Oncology*, 16, 1227.
20. Zhou, H., Li, M., & Chen, W. (2024). Reinforcement learning applications in surgical training systems. *IEEE Transactions on Medical Robotics*, 8(3), 189-203.
21. Thompson, K., Williams, D., & Jones, P. (2024). Content generation capabilities in advanced language models for medical education. *Medical Education Technology*, 13(1), 67-81.
22. Catalano, E. (2024). Silvia Tornusciolo Mercatorum University, Italy Enzapaola Catalano Pegaso University, Italy Tonia De Giuseppe. *Disruptive Technologies in Education and Workforce Development*, 251.
23. Martinez, C., Lee, K., & Park, J. (2024). Limitations of multiple-choice assessment in evaluating clinical reasoning for minimally invasive surgery. *Surgical Education Assessment*, 8(4), 201-215.
24. Foster, N., Adams, B., & Clark, H. (2024). Advanced evaluation metrics for medical AI systems in specialized contexts. *AI Evaluation Methods*, 12(1), 34-49.

Copyright: ©2025 Bevan L. Staple, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.