Situational context embedding and knowledge-based grounding for LLMs in surgical training assistance

Ali Bahari Malayeri[1], Matthias Seibold[1], Felix Öttl[2], Nicola Cavalcanti[1], Tatiana Gerth[3], Roni Lekar[3], Armando Hoch[2], Patrick Zingg[2], and Philipp Fürnstahl[1]

[1] Research in Orthopedic Computer Science (ROCS), Balgrist University Hospital, University of Zurich, Forchstrasse 340, 8008 Zürich, Switzerland
ali.baharimalayeri@balgrist.ch
[2] Department of Orthopedics, Balgrist University Hospital, University of Zurich, Forchstrasse 340, 8008 Zürich, Switzerland
[3] Institute of Data Science (IDS), School of Engineering, Zurich University of Applied Sciences (ZHAW), Technikumstrasse 81, 8400 Winterthur, Switzerland

Abstract. High-quality surgical training requires timely, context-aware feedback, yet expert supervision is resource intensive. Surgical simulators enable scalable, self-paced practice, but trainees often lack an instructor for in-the-moment questions. Although large language models can provide natural-language guidance, they are prone to hallucinations, generic recommendations, and insufficient grounding in the concrete simulator context. We present a proof-of-concept of a LLM-based tutoring system for Total Hip Arthroplasty (THA) training that explicitly grounds responses in both simulator-derived context and curated domain knowledge. Situational context is derived directly from the training simulator, integrating workflow phase information, procedural targets, and spatial relationships between anatomical structures, instruments, and implants. These structured state representations are transformed into anatomically meaningful clinical descriptions to enable context-aware tutoring. In parallel, the system incorporates simulator-specific documentation and expert THA textbook sources incorporated through retrieval-augmented generation (RAG). To mitigate unsupported outputs, our approach introduces an evidence-first policy with calibrated abstention when evidence is insufficient. We evaluate the proposed approach in a blinded expert study with 60 THA-related questions in pre-operative, process-related, and situational categories. Five senior THA surgeons rate anonymized responses from the proposed system, a baseline LLM, and expert reference answers. The results indicate that the proposed framework effectively integrates situational context and medical knowledge and achieves a significant improvement of the response quality in comparison to a general-purpose LLM baseline. However, human expert feedback remains vital for effective surgical training guidance.

Keywords: Surgical Training · Question Answering · Large Language Model

1 Introduction

High-quality surgical treatment relies fundamentally on rigorous medical education and structured surgical training [28]. While traditional apprenticeship-based paradigms, such as the "see one, do one, teach one" paradigm, emphasize direct expert supervision and feedback-driven learning, modern healthcare systems face increasing pressure due to staff shortages, rising costs, and higher case volumes, substantially constraining the time for supervised training [13, 15]. To mitigate these constraints, simulation-based surgical training has evolved into a central component of modern curricula, spanning low-fidelity physical trainers (e.g., box trainers for basic laparoscopic skills), immersive virtual-reality simulators, and augmented/extended-reality systems that overlay procedural guidance onto the trainee's view. These components enable standardized, structured, self-paced, and scalable psychomotor practice [21, 17, 26, 29].

However, simulation-based training does not inherently resolve the underlying shortage of expert supervision; rather, it partially externalizes hands-on instruction from the operating room to dedicated training environments. Critical aspects of surgical training such as real-time clarifications, contextual explanation of anatomical variations, and individualized feedback, remain difficult to provide without instructor involvement.

Fully self-directed simulator training in open surgery is therefore still limited by the absence of interactive, context-aware instructional support. Trainees are unable to ask questions, test alternative strategies, or receive real-time feedback in the absence of an expert. Given the fact that senior surgeons' time is a major cost driver, this reliance becomes a structural bottleneck. LLMs have shown promise for automated tutoring systems [25, 5], and recent literature demonstrates their successful application in broad medical education and clinical reasoning, with evaluations confirming their ability to answer complex diagnostic queries and even pass licensure exams [4, 14]. In the surgical domain, initial studies suggest that LLMs can assist with procedural knowledge and high-level decision support. Beyond language interaction, multimodal LLMs have been explored for surgical visual question answering (VQA), scene understanding, and action planning based on recorded operative data. These works demonstrate that models can interpret surgical situations [18–20] and therefore potentially bridge the instructional gap in self-directed training.

However, current approaches remain limited in several aspects. They are predominantly retrospective and do not address simulator-grounded tutoring that requires explicit spatial context, tight workflow integration, and near real-time responsiveness. When used in real-world tasks, LLMs still suffer from hallucinations which in surgical training may mislead trainees and compromise training outcomes [7, 8, 2, 1]. These limitations highlight two underlying challenges in LLM-based tutoring approaches: (1) the lack of explicit situation-aware and spatial context that couples semantic reasoning with simulator states (e.g., phase, active tool, and pose deviations), and (2) insufficient safeguarding mechanisms to constrain model output to evidence-based content [24, 30].

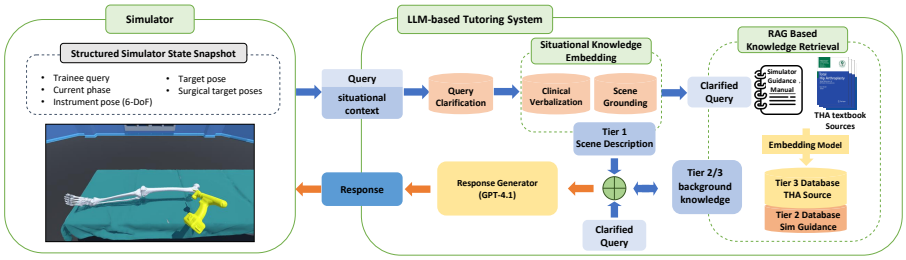


Fig. 1. Overview of the proposed LLM-based Tutoring System.

In this work, we propose a novel concept for LLM-based tutoring systems that integrates situational context and knowledge-based grounding. The contributions of our work can be summarized as follows:

- A hierarchical knowledge-integration framework that structures tutoring information into simulator-derived situational context and two curated document tiers integrated through RAG and calibrated abstention.
- Integration of simulator context via deterministic processing of phase-dependent serialized snapshots, producing anatomically meaningful spatial descriptors for situation-aware feedback.
- A thorough evaluation in a blinded expert study with five senior Total Hip Replacement (THA) surgeons using an expert-defined question catalog consisting of pre-operative, process-related, and situation-aware categories to benchmark the proposed framework against expert references and baseline LLM.

2 Materials and Methods

In the following section, we first describe the overall architecture of the proposed LLM-based tutoring system (Fig. 1), detailing its main components and implementation. Subsequently, we describe the design of the expert user study conducted for the evaluation of the proposed system.

2.1 Architecture

Each request to our system follows an input–processing–output flow. The simulator provides (i) the trainee's query and (ii) a phase-dependent snapshot of the simulator context. The tutoring framework then combines this simulator-derived context with curated background knowledge and returns a response to the simulator.

Upon receiving a user question, the system first applies a *query clarification* module. This step is necessary because trainees questions are often underspecified, ambiguous or lack procedural context which can lead to suboptimal retrieval

or imprecise responses. The module reformulates the input into a context-aware, task-specific representation while preserving the original intent and scope (e.g., "Which comorbidities can occur?" → "Which comorbidities should be monitored in patients undergoing total hip arthroplasty (THA)?"). The clarified question is only used for downstream retrieval and does not constitute the final answer.

We structure the system's processing pipeline into three tiers. Tier 1 is generated from structured simulator context through a *situational knowledge embedding* component (Section 2.2), which transforms scene grounding into concise, anatomically meaningful clinical descriptions. In parallel, the system leverages RAG to process Tier 2 and 3 medical background knowledge (Section 2.3). For all queries, the clarified question, the Tier 1 scene description, and the Tier 2/3 background knowledge are fed to the *response generator* LLM, which is constrained to generate answers only on the basis of the supplied evidence. Text retrieval uses a back-off strategy with calibrated abstention. In the following sections, we will introduce the above-described components in detail.

2.2 Situational Knowledge Embedding

The simulator-derived context is formally defined as the *situational context*, which comprises the current workflow phase, the active instrument, its 6-DoF pose, a set of anatomical landmark points, and phase-specific surgical targets describing the intended anatomy–instrument or anatomy–implant configuration for the current phase. Each phase may include one or more surgical targets, depending on the training objective. To demonstrate our concept, we implement a Unity-based THA training application following the surgical steps defined in [29], with surgical targets such as the femoral neck osteotomy plane, acetabular reaming alignment, planned acetabular cup pose, and the femoral broach or stem pose. Poses and targets are parameterized in the simulator coordinate frame using translations in metric units and rotations as Euler angles. The parameters describing the current simulator state are sent to the LLM-based tutoring system alongside the user question. A phase-specific *scene grounding* module deterministically computes deviation from the current phase-specific surgical targets: translational error is reported as component-wise differences and Euclidean distance, and rotational error as per-axis Euler-angle differences. All computations are performed outside the LLM to prevent arithmetic inconsistencies and ensure reproducible feedback. The *clinical verbalization* module maps translational and rotational deviation metrics into structured text with a concise clinical terminology (*Tier 1 Scene Description*) derived from the anatomical coordinate system with anterior–posterior, medial–lateral, and proximal–distal axes. This approach enables the translation of quantitative data into instruction-style feedback that can be readily interpreted by trainees. The system handles deviations following threshold-based rules. If translational and rotational deviations remain within predefined, phase-specific tolerance thresholds, the system provides guidance in medical terminology specifying the required positional and angular adjustments relative to the surgical target. If deviations exceed these thresholds, instant guid-

ance is suppressed and a reinitialization directive is issued, instructing the trainee to restart the current procedural step.

2.3 RAG-based Knowledge Retrieval

To systematically incorporate curated background knowledge, we introduce two tiers of medical reference documents that are processed through a RAG pipeline. Tier 2 comprises custom training simulator-specific guidance documents including descriptions of workflow phases, usage of surgical instruments, surgical terminology, and common pitfalls. Tier 3 constitutes our Background Corpus and contains comprehensive THA knowledge from expert-recommended textbooks for orthopedic surgery [10–12, 23, 3]. This hierarchical organization follows recent work on routed retrieval in multi-corpus settings, where restricting retrieval to the most relevant source reduces distracting or conflicting context [31, 6].

The documents are segmented into overlapping chunks (4000 characters, 200 overlap) and encoded into dense vector embeddings using a pretrained text embedding model. For efficient nearest-neighbor retrieval, embeddings are stored in a FAISS index supporting L2-distance search [9]. At inference time, the clarified query is embedded and used to retrieve the top-k chunks ($k = 20$). Retrieved candidates are filtered by an L2-distance threshold τ ($\tau = 0.5$; lower distance indicates higher similarity), and the retained excerpts are passed to the *response generator* for evidence-grounded response generation.

The system first queries Tier 2. If no retrieved chunk satisfies the similarity threshold, or if the model indicates insufficient supporting evidence, retrieval is escalated to Tier 3 using identical parameters. If evidence is still insufficient, the system abstains from generating a response. In this case, the system emits an explicit non-answer rather than producing ungrounded text, ensuring that responses are only generated when sufficient supporting evidence is available.

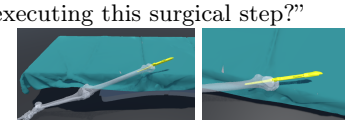
2.4 Implementation Details

The proposed system is implemented as a lightweight external service that interfaces with the Unity simulator via an HTTP request–response protocol. The *response generator* component is based on `gpt-4.1` with temperature set to 0 to ensure deterministic behaviour. The *query clarification* module uses `gpt-4o` also with temperature 0 and is employed only to reformulate the input query for retrieval. All methods are implemented in Unity version 6000.0.23f1 and Python version 3.11.9. The source code, the question catalog, and the expert reference answers are publicly available at https://github.com/ali-b-m/surgical-llm-tutor-tha. Our framework is LLM-agnostic; the proprietary-API dependency affects only the underlying language model and can be substituted with an open model.

2.5 Expert User Study

We conducted a blinded expert study comparing three answer sources: (i) Proposed system, (ii) baseline `gpt-4.1` queried via the standard API (vision-language

Table 1. Question categories with representative examples spanning pre-operative, process-related and situational.

Category	Example 1	Example 2
Pre-operative	"Which image modalities do we need for the surgical planning?"	"When and how do we need to adapt for differences in leg length?"
Process-related	"What are the different types of wound closure techniques used in THA?"	"How can you intraoperatively assess for impingement?"
Situational (simulator)	"Am I in the correct position for executing this surgical step?" 	"How should I position the patient for this phase?" 

input for situational questions), and (iii) Consensus Expert Reference answers. The baseline (ii) received the user question along with a brief role instruction (THA simulator tutor with concise answer constraint). For situational questions, it additionally received two screenshots from the simulator (near and far view) as contextual input. The evaluation set comprised 60 questions equally distributed across three categories: pre-operative, process-related, and situational questions requiring simulator context (Table 1). Questions and Consensus Expert Reference answers were created by a clinical authoring group consisting of two experienced THA surgeons and two residents. For each question, one answer per system was generated, yielding three answer sets.

In the evaluation, five independent, previously uninvolved senior THA surgeons participated as evaluators. A custom graphical user interface presented each question with the three blinded answers in randomized order. For situational questions, the corresponding simulator screenshots were shown alongside to provide visual context. For each displayed answer, evaluators rated five criteria namely accuracy, relevance, comprehensiveness, clarity, and trust using a 5-point Likert scale adapted from QUEST [27], a common evaluation benchmark for LLMs in medical question answering.

3 Results

Fig. 2 summarizes the results from the expert ratings across 60 questions (20 per category) and five THA surgeons for the three answer sources: LLM-based Tutoring System, baseline GPT-4.1, and Expert Reference. The mean overall scores (mean $\pm$ SD across $n = 60$ questions) were 4.424 ± 0.469 for the expert reference, 3.669 ± 0.754 for our system, and 3.122 ± 0.846 for the baseline model. Pairwise statistical significance between answer sources was assessed at $\alpha = 0.05$ using a two-sided Wilcoxon signed-rank test. The proposed system significantly outperformed the baseline LLM (all $p < 1.2 \times 10^{-4}$), while expert reference answers scored significantly higher than the proposed system (all $p < 1.8 \times$

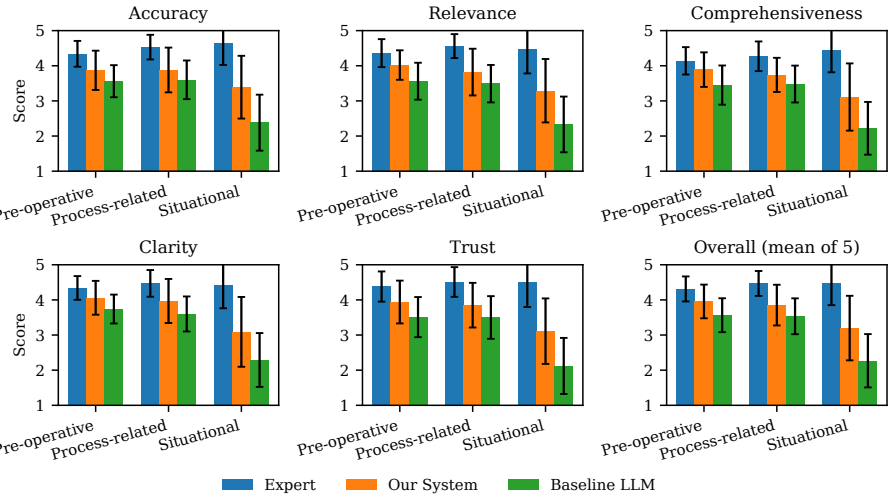


Fig. 2. Expert study ratings by question category and source. Bars show mean Likert scores; error bars show ±1 SD across questions.

Table 2. Reference-based text similarity to the Expert Reference answers across 60 questions, reported as median with 95% bootstrap confidence intervals in brackets.

System	ROUGE-1 (Precision)	SBERT cosine
Baseline LLM	0.276 [0.261,0.310]	0.656 [0.609,0.707]
LLM-based Tutoring System (Ours)	0.320 [0.298,0.347]	0.695 [0.655,0.734]
w/o Query Clarification	0.323 [0.277,0.364]	0.664 [0.623,0.698]
w/o Tier 1	0.326 [0.294,0.377]	0.651 [0.635,0.699]
w/o Tier 2	0.315 [0.298,0.351]	0.671 [0.628,0.720]
w/o Tier 3	0.000 [0.000,0.086]	-0.005 [-0.029,0.035]

10^{-6}). Inter-rater reliability was *good* across criteria (ICC(2,k)=0.786–0.834), supporting that the observed differences are not due to noisy ratings.

Ablation Study | We evaluate the contribution of individual components using reference-based text similarity to the Expert Reference answers, reporting ROUGE-1 [16] and Sentence-BERT cosine similarity [22]. Table 2 summarizes results across all 60 questions (median with 95% bootstrap confidence intervals).

4 Discussion

The results of our expert user study with 5 experienced hip surgeons demonstrate that, while human expert feedback remains the gold standard for simulator-based surgical training, structured grounding substantially enhances the effectiveness of LLM-based tutoring systems. In particular, the proposed combination of situational context embedding and knowledge-based grounding outperformed

a general-purpose LLM baseline across all criteria. The results in Fig. 2 show that the answers of our proposed system were consistently preferred over the baseline. While the baseline LLM produced credible answers, we show that grounding the responses in a curated database of medical and domain-specific background knowledge further improves the perceived quality of the answers. Largest performance gains were observed for situational questions requiring context-dependent guidance, highlighting the importance of structured scene grounding in interactive training environments.

To analyze individual components, we conducted an ablation study using automatic LLM evaluation metrics, as a full expert-based ablation was not feasible given the required clinical resources. While these metrics allow controlled comparisons, they may not be sufficiently sensitive to capture nuanced differences in clinical meaningfulness, which likely explains the modest changes observed across configurations. Removing Tier 3 yields ROUGE-1 $\approx$ 0 and SBERT $\approx$ 0 by design: without textbook-derived evidence the system abstains rather than producing ungrounded text, so the near-zero similarity reflects the intended safety mechanism rather than a system collapse, highlighting the importance of curated medical background knowledge in the proposed framework. The baseline (GPT-4.1 with a brief role instruction) represents the most prevalent real-world alternative, and the leave-one-component-out variants in Table 2 directly capture the LLM-only versus LLM+RAG performance gap across all components.

Our evaluation was based on an expert-defined question catalog and did not assess longitudinal effects on trainee learning outcomes. However, the substantial clinical resources required for this study must be considered when assessing the feasibility of more extensive expert-based evaluations. Two senior, board-certified and two resident surgeons from our university hospital were involved in the definition of the questionnaire, and 5 board-certified surgeons spent more than two hours each in the expert study. The proposed system relied on external model APIs, which may have introduced variability due to provider-side updates or service conditions, potentially affecting response consistency even when inputs remained unchanged. In our study, we rely on OpenAI's model family and base our system on `gpt-4.1` which was one of the most promising models at the time of project definition. As the landscape of LLMs is changing rapidly, models of different providers and latest model versions should be tested in future research. Furthermore, the current spatial mapping is embedded within a software-only Unity-based surgical training simulation, as illustrated in Fig. 1. Our core modules (the RAG pipeline, the abstention mechanism, and query clarification) are procedure-agnostic by design. Adapting the system to a new procedure or simulator primarily requires swapping the curated knowledge base and redefining the simulator-state representation, which are largely engineering tasks. Demonstrating this transfer is a concrete hypothesis for future work.

5 Conclusion

In this work, we introduce a LLM-based system that combines explicit situational context and hierarchical, knowledge-based grounding for effective ques-

tion answering in surgical training simulators. Our study evaluates system-level response quality and does not assess longitudinal training effectiveness or surgical-performance outcomes, which require a controlled longitudinal trial. Our results show that, although human experts remain indispensable, the proposed system can be useful for self-paced surgical training assistance by delivering context-aware, anatomically grounded feedback and significantly surpasses general-purpose LLMs in expert evaluation.

Disclosure of Interests. The authors report there are no competing interests to declare.

References

1. Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J.A., Pimenta, D.: A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. npj Digital Medicine **8**(1), 274 (2025)
2. Azamfirei, R., Kudchadkar, S.R., Fackler, J.: Large language models and the perils of their hallucinations. Critical Care **27**(1), 120 (2023)
3. Bagaria, V. (ed.): Total Hip Replacement: An Overview. IntechOpen, London (2018). https://doi.org/10.5772/intechopen.71398
4. Bicknell, B.T., Butler, D., Whalen, S., Ricks, J., Dixon, C.J., Clark, A.B., Spaedy, O., Skelton, A., Edupuganti, N., Dzubinski, L., et al.: Chatgpt-4 omni performance in usmle disciplines and clinical skills: comparative analysis. JMIR Medical Education **10**(1), e63430 (2024)
5. Holderried, F., Stegemann-Philipps, C., Herrmann-Werner, A., Festl-Wietek, T., Holderried, M., Eickhoff, C., Mahling, M., et al.: A language model–powered simulated patient with automated feedback for history taking: Prospective study. JMIR Medical Education **10**(1), e59213 (2024)
6. Huang, H., Huang, Y., Yang, J., Pan, Z., Chen, Y., Ma, K., Chen, H., Cheng, J.: Retrieval-augmented generation with hierarchical knowledge. arXiv preprint arXiv:2503.10150 (2025)
7. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2), 1–55 (2025)
8. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM computing surveys **55**(12), 1–38 (2023)
9. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. IEEE transactions on big data **7**(3), 535–547 (2019)
10. Knahr, K. (ed.): Tribology in Total Hip Arthroplasty. Springer, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19429-0
11. Knahr, K. (ed.): Total Hip Arthroplasty: Wear Behaviour of Different Articulations. Springer, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-27361-2
12. Knahr, K. (ed.): Total Hip Arthroplasty: Tribological Considerations and Clinical Consequences. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-35653-7
13. Kotsis, S.V., Chung, K.C.: Application of the "see one, do one, teach one" concept in surgical training. Plastic and reconstructive surgery **131**(5), 1194–1201 (2013)
14. Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., et al.: Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. PLoS digital health **2**(2), e0000198 (2023)
15. Lauer, C.I., Shabahang, M.M., Restivo, B., Lane, S., Hayek, S., Dove, J., Ellison, H.B., Pica, E., Ryer, E.J.: The value of surgical graduate medical education (gme) programs within an integrated health care system. Journal of surgical education **76**(6), e173–e181 (2019)
16. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)

17. Mao, R.Q., Lan, L., Kay, J., Lohre, R., Ayeni, O.R., Goel, D.P., et al.: Immersive virtual reality for surgical training: a systematic review. Journal of Surgical Research **268**, 40–58 (2021)
18. Ong, C.S., Obey, N.T., Zheng, Y., Cohan, A., Schneider, E.B.: Surgeryllm: a retrieval-augmented generation large language model framework for surgical decision support and workflow enhancement. npj Digital Medicine **7**(1), 364 (2024)
19. Palenzuela, D.L., Mullen, J.T., Phitayakorn, R.: Ai versus md: Evaluating the surgical decision-making accuracy of chatgpt-4. Surgery **176**(2), 241–245 (2024)
20. Pressman, S.M., Borna, S., Gomez-Cabello, C.A., Haider, S.A., Haider, C.R., Forte, A.J.: Clinical and surgical applications of large language models: a systematic review. Journal of Clinical Medicine **13**(11), 3041 (2024)
21. Raj, D.R.S., Kumar, S., Nallathamby, K., Raj, K., Hristova, M.: A systematic review of the learning curves of novices and trainees to achieve proficiency in laparoscopic skills: virtual reality simulator versus box trainer. Cureus **16**(11) (2024)
22. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
23. Sharma, M. (ed.): Hip Arthroplasty: Current and Future Directions. Springer, Singapore (2023). https://doi.org/10.1007/978-981-99-5517-6
24. Shen, Y., Li, C., Liu, B., Li, C.Y., Porras, T., Unberath, M.: Operating Room Workflow Analysis via Reasoning Segmentation over Digital Twins . In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15968, pp. 415 – 424. Springer Nature Switzerland (October 2025)
25. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al.: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (2023)
26. Suresh, D., Aydin, A., James, S., Ahmed, K., Dasgupta, P.: The role of augmented reality in surgical training: a systematic review. Surgical Innovation **30**(3), 366–382 (2023)
27. Tam, T.Y.C., Sivarajkumar, S., Kapoor, S., Stolyar, A.V., Polanska, K., McCarthy, K.R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., et al.: A framework for human evaluation of large language models in healthcare derived from literature review. NPJ digital medicine **7**(1), 258 (2024)
28. Theodoulou, I., Sideris, M., Lawal, K., Nicolaides, M., Dedeilia, A., Emin, E.I., Tsoulfas, G., Papalois, V., Velmahos, G., Papalois, A.: Retrospective qualitative study evaluating the application of ig4 curriculum: an adaptable concept for holistic surgical education. BMJ open **10**(2), e033181 (2020)
29. Wu, L., Seibold, M., Cavalcanti, N.A., Hein, J., Gerth, T., Lekar, R., Hoch, A., Vlachopoulos, L., Grabner, H., Zingg, P., et al.: A novel augmented reality-based simulator for enhancing orthopedic surgical training. Computers in Biology and Medicine **185**, 109536 (2025)
30. Xu, M., Huang, Z., Zhang, J., Zhang, X., Dou, Q.: Surgical Action Planning with Large Language Models . In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15968, pp. 563 – 572. Springer Nature Switzerland (October 2025)
31. Zhong, Z., Liu, H., Cui, X., Zhang, X., Qin, Z.: Mix-of-granularity: Optimize the chunking granularity for retrieval-augmented generation. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 5756–5774 (2025)