BMC Musculoskeletal Disorders

RESEARCH

Open Access

Evaluating the performance of five large language models in answering Delphi consensus questions relating to patellar instability and medial patellofemoral ligament reconstruction

Prushoth Vivekanantha[1*], Dan Cohen[1], David Slawaska-Eng[1], Kanto Nagai[2], Magdalena Tarchala[3], Bogdan Matache[3], Laurie Hiemstra[4], Robert Longstaffe[5], Bryson Lesniak[6], Amit Meena[7], Sachin Tapasvi[8], Petri Sillanpää[9], Patrick Grzela[1], Daniel Lamanna[1], Kristian Samuelsson[10,11*] and Darren de SA[1]

Abstract

Purpose Artificial intelligence (AI) has become incredibly popular over the past several years, with large language models (LLMs) offering the possibility of revolutionizing the way healthcare information is shared with patients. However, to prevent the spread of misinformation, analyzing the accuracy of answers from these LLMs is essential. This study will aim to assess the accuracy of five freely accessible chatbots by specifically evaluating their responses to questions about patellofemoral instability (PFI). The secondary objective will be to compare the different chatbots, to distinguish which LLM offers the most accurate set of responses.

Methods Ten questions were selected from a previously published international Delphi Consensus study pertaining to patellar instability, and posed to ChatGPT4o, Perplexity AI, Bing CoPilot, Claude2, and Google Gemini. Responses were assessed for accuracy using the validated Mika score by eight Orthopedic surgeons who have completed fellowship training in sports-medicine. Median responses amongst the eight reviewers for each question were compared using the Kruskal-Wallis and Dunn's post-hoc tests. Percentages of each Mika score distribution were compared using Pearson's chi-square test. P-values less than or equal to 0.05 were considered significant. The Gwet's AC2 coefficient was calculated to assess for inter-rater agreement, corrected for chance and employing quadratic weights.

Results ChatGPT4o and Claude2 had the highest percentage of reviews (38/80, 47.5%) considered to be an "excellent response not requiring classification", or a Mika score of 1. Google Gemini had the highest percentage of reviews (17/80, 21.3%) considered to be "unsatisfactory requiring substantial clarification", or a Mika score of 4 ($p < 0.001$).

*Correspondence:
Prushoth Vivekanantha
prushoth.vivekanantha@medportal.ca
Kristian Samuelsson
kristian@samuelsson.cc; Kristian.samuelsson@gu.se

Full list of author information is available at the end of the article

The median ± interquartile range (IQR) Mika scores was 2 (1) for ChatGPT4o and Perplexity AI, 2 (2) for Bing CoPilot and Claude2, and 3 (2) for Google Gemini. Median responses were not significantly different between ChatGPT4o, Perplexity AI, Bing CoPilot, and Claude2, however all four statistically outperformed Google Gemini ($p < 0.05$). Inter-rater agreement was classified as moderate ($0.40 > AC2 \geq 0.60$) for ChatGPT, Perplexity AI, Bing CoPilot, and Claude2, while there was no agreement for Google Gemini ($AC2 < 0$).

Conclusion Current free access LLMs (ChatGPT4o, Perplexity AI, Bing CoPilot, and Claude2) predominantly provide satisfactory responses requiring minimal clarification to standardized questions relating to patellar instability. Google Gemini statistically underperformed in accuracy relative to the other four LLMs, with most answers requiring moderate clarification. Furthermore, inter-rater agreement was moderate for all LLMs apart from Google Gemini, which had no agreement. These findings advocate for the utility of existing LLMs in serving as an adjunct to physicians and surgeons in providing patients information pertaining to patellar instability.

Level of evidence: V Keywords Large language models, Chatbots, Artificial intelligence, Patellar instability, Medial patellofemoral ligament reconstruction

Introduction

Artificial intelligence (AI) has become increasingly popular over the past few years [1]. One of the most mainstream usages of AI is in the form of large language models (LLMs), such as ChatGPT (Open AI, San Francisco, United States) or Perplexity (Perplexity AI, San Francisco, United States) [2, 3]. These LLMs offer novel opportunities to enhance patient education and care [4–6]. Recent studies have investigated the quality of responses from LLMs, notably ChatGPT, to various questions pertaining to orthopedic pathologies and surgeries. This includes anterior cruciate ligament reconstruction (ACLR), hip arthroscopy, and ulnar collateral ligament reconstruction (UCLR), generally finding satisfactory responses when reviewed by experts [7–9].

Patellofemoral instability (PFI) is a common orthopedic condition that may require surgical intervention in the form of medial patellofemoral ligament reconstruction (MPFLR) [10, 11]. It has an incidence between 21.6 and 49.7 per 100 000 persons [12]. PFI has been shown to cause similar levels of dysfunction to the knee as that of an ACL tear [13]. It is the responsibility of the healthcare team to ensure that patients receive accurate, digestible, and up-to-date information about their condition, and what may be required from a treatment and recovery point of view. However, LLMs may serve as a useful adjunct, where patients would be able to have immediate access to information outside of hospital and clinic visits. Given this, assessing the use of LLMs responses for pathologies such as PFI is essential to ensure patient safety.

This study will aim to assess the accuracy of five freely accessible chatbots: ChatGPT4o, Perplexity AI, Claude2 (Anthropic, San Francisco, United States), Microsoft Copilot (Microsoft, Redmond, Washington), and Gemini (Google DeepMind, London, United Kingdom), by specifically evaluating their responses to questions about PFI. The secondary objective will be to compare the different chatbots, to distinguish which LLM offers the most accurate set of responses.

Materials and methods

Question selection

Questions that were selected were based upon a previous modified Delphi Consensus statement study on patellar instability [13, 14]. This consensus statement was developed in collaboration with 60 surgeons, all either members of one or more of the following societies: American Orthopedic Society for Sports Medicine (AOSSM), Arthroscopy Association of North America (AANA), European Society of Sports Traumatology, Knee Surgery and Arthroscopy, International Society of Arthroscopy (ESSKA), Knee Surgery and Sports Medicine (ISAKOS), and the Patellofemoral Foundation [13, 14]. Questions were only chosen if there was unanimous consensus or if there was strong consensus, as per the original manuscript. The questions were as follows:

1. What factors of patient history should be evaluated in the setting of patellar instability?
2. What aspects of the physical exam should be performed in patients with patellar instability?
3. When should advanced imaging (MRI/CT) be performed in patients presenting with patellar instability?
4. When should patients start range of motion exercises when undergoing non-operative management for patellar instability?
5. What are the indications for nonoperative management for patients with patellar instability?
6. What are the contraindications for nonoperative management for patients with patellar instability?
7. What are the indications for MPFL reconstruction for patients with patellar instability?

8. Are there any different considerations that should be made in pediatric patients undergoing MPFL reconstruction for patellar instability?
9. What clinical factors may influence your decision to perform a TTO for patellar instability?
10. Which factor(s), if any, should be considered when deciding which trochleoplasty technique to perform?

Accuracy assessment of responses

The ten questions were inputted into ChatGPT4o, Perplexity AI, Claude2, Microsoft Copilot, and Google Gemini in August 2024. There were no prior prompts used nor were there any repeat or follow-up questions asked. Answers were assessed by eight orthopedic surgeons with fellowship training in sports medicine using the rating system described by Mika et al. [15, 16]. This grading system has been used extensively in the past by a variety of studies with similar aims [8, 17, 18]. This system consists of four categories: Response accuracy was rated as (1) "excellent response not requiring clarification" if it provided fundamentally factual information free of inaccuracies; (2) "satisfactory requiring minimal clarification" if it provided correct information but was missing some finer points or nuances; (3) "satisfactory requiring moderate clarification" if it provided outdated information or information that was not relevant to the question asked; or (4) "unsatisfactory requiring substantial clarification" if it provided incorrect or overly generalized information that could be conceivably misinterpreted or detrimental [15, 16]. Surgeons were not blinded during this study.

Statistical analysis

Descriptive statistics, such as medians, interquartile ranges (IQR), means, ranges, and percentages were utilized. Median scores for each LLM were calculated, compared using the nonparametric Kruskal-Wallis and Dunn's post-hoc tests. A two-tailed Pearson's chi-square test was conducted comparing percentages of Mika scores, with adjusted standardized residuals used to identify where significant differences occurred. Values greater than positive or negative 2 were considered percentages differing significantly from expectation. P-values less than or equal to 0.05 were considered to be statistically significant. Gwet's AC2 coefficient employing quadratic weights was utilized for assessment of agreement. This coefficient accommodates weighted values for ordinal variables, making it suitable for comparing Mika score values. The Landis and Koch categorization was adopted as per previous studies, given the lack of well-established thresholds for Gwet's AC2 coefficient [19–21]. The categorization of the AC2 coefficient was defined a priori as: 1.00 >AC2 ≥ 0.80 indicates almost perfect agreement, 0.80 >AC2 ≥ 0.60 indicates substantial agreement, 0.60 >AC2 ≥ 0.40 indicates moderate agreement, 0.40 >AC2 ≥ 0.60 indicates fair agreement, 0.20 >AC2 ≥ 0.00 indicates slight agreement and a AC2 score = 0 indicates no agreement [20]. For each Gwet's AC2 coefficient, 95% confidence intervals (CI) were also calculated. All statistics were performed using Excel (Microsoft, Redmond, Washington, USA) or Python (version 3.11; Python Software Foundation, Wilmington, Delaware, USA).

Results

Accuracy of LLM responses

The median (IQR) Mika score for ChatGPT4o and Perplexity AI was 2 (1) for both, while the median (IQR) Mika score for Bing CoPilot and Claude2 was 2 (2) for both. The median (IQR) Mika score for Google Gemini was 3 (2) (Table 1). Amongst 80 reviews to responses, ChatGPT4o and Claude2 had the highest amount rated as a Mika 1 ($n = 38$; 47.5%), followed by Perplexity AI ($n = 33$; 41.3%). Google Gemini had 31.3% ($n = 25$) and 21.3% ($n = 17$) rated as a Mika 3 and 4, respectively (Fig. 1). Statistical comparisons using the Kruskal-Wallis test found a significant difference between LLMs ($p = 0.0005$). Dunn's post-hoc tests found that ChatGPT ($p = 0.00002$), Perplexity AI ($p = 0.002$), Bing Copilot ($p = 0.01$), Claude2 ($p = 0.0002$) all outperformed Google Gemini in accuracy. No significant differences were found between ChatGPT, Perplexity AI, Bing CoPilot, and Claude2. Statistical comparisons of distribution of Mika scores across the fives LLMs using the Pearson's chi-square test was also significant ($X^2 = 48.74$, $p < 0.001$). Using adjusted standardized residuals, Google Gemini was found to have significantly fewer Mika scores of 1 (adjusted residual = −2.41) and 2 (adjusted residual = −2.13), but more Mika scores of 4 (adjusted residual = + 5.40). Claude2 was found to have significantly fewer Mika scores of 4 (adjusted residual = −2.31). Additionally, ChatGPT4o had significantly fewer Mika scores of 3 (adjusted residual = −2.35) while Bing CoPilot had significantly more scores of 2 (adjusted residual = + 1.97).

All chatbots had at least one response that was rated as "unsatisfactory requiring substantial clarification", or a Mika score of 4, with all having at least one rating of 4 for question 4 ("*When should patients start range of motion exercises when undergoing non-operative management for patellar instability?*"). For ChatGPT4o, there were five total responses rated as a 4, however they were all for different questions (#4, #5, #7, #9, #10), with three (60%) from one reviewer. For Perplexity AI, there were three total responses rated as a 4, for two different questions (#3, #4) amongst two different reviewers. Bing CoPilot also had three total responses rated as a 4, amongst three different questions (#4, #6, #8) and three different reviewers. Claude2 only had one response rated as a 4 (#4). Five reviewers rated at least one response as a 4 for Google Gemini amongst nine of 10 total questions (all but #1).

Table 1 Individual reviewer responses to answers by large language models (LLMs)

ChatGPT4o								
Question#	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6	Reviewer 7	Reviewer 8
1	1	1	2	2	3	2	1	1
2	2	2	2	3	3	1	2	1
3	2	1	1	3	3	1	1	1
4	3	1	2	3	4	1	2	1
5	3	2	1	4	1	1	1	1
6	2	2	1	2	1	1	1	1
7	4	2	1	2	2	1	2	1
8	3	1	2	2	2	1	1	1
9	3	1	2	4	2	1	2	1
10	3	1	2	4	2	1	1	1
Perplexity AI								
Question #	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6	Reviewer 7	Reviewer 8
1	3	3	2	1	3	2	3	1
2	3	1	2	1	2	1	2	1
3	1	2	2	1	4	4	3	1
4	2	1	2	1	4	1	2	1
5	2	2	2	1	2	1	2	1
6	2	1	2	2	3	1	2	1
7	2	1	3	2	2	1	3	1
8	2	1	2	2	3	1	3	1
9	1	1	2	2	3	1	3	1
10	3	1	2	1	3	1	3	1
Bing CoPilot								
Question #	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6	Reviewer 7	Reviewer 8
1	3	2	2	2	2	2	2	1
2	3	2	3	2	2	1	2	1
3	2	1	2	2	3	1	3	1
4	1	1	1	2	4	1	3	1
5	3	2	2	1	2	1	3	1
6	3	4	3	2	1	2	3	1
7	2	3	3	2	3	1	2	1
8	3	2	2	2	3	1	4	1
9	2	1	3	2	3	1	3	1
10	2	1	2	2	3	1	2	1
Claude2								
Question #	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6	Reviewer 7	Reviewer 8
1	1	1	1	1	3	1	2	1
2	2	1	1	1	3	1	3	1
3	3	1	1	1	3	1	3	1
4	1	1	2	3	4	1	3	1
5	3	2	2	1	2	1	2	1
6	2	1	2	3	2	1	3	1
7	3	1	2	3	2	1	2	1
8	3	2	3	3	2	1	2	1
9	1	1	3	3	3	1	2	1
10	2	1	3	3	3	1	3	1
Google Gemini								
Question #	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6	Reviewer 7	Reviewer 8
1	3	2	3	3	2	1	1	1
2	4	4	3	3	1	1	3	1
3	4	3	3	4	2	1	2	1
4	2	1	3	4	4	1	3	1

Table 1 (continued)

5	3	4	3	4	1	2	3	1
6	4	3	3	2	2	2	3	1
7	4	3	3	3	2	1	3	1
8	4	4	3	3	2	3	2	1
9	4	4	3	4	1	4	1	1
10	2	2	3	4	2	1	2	1

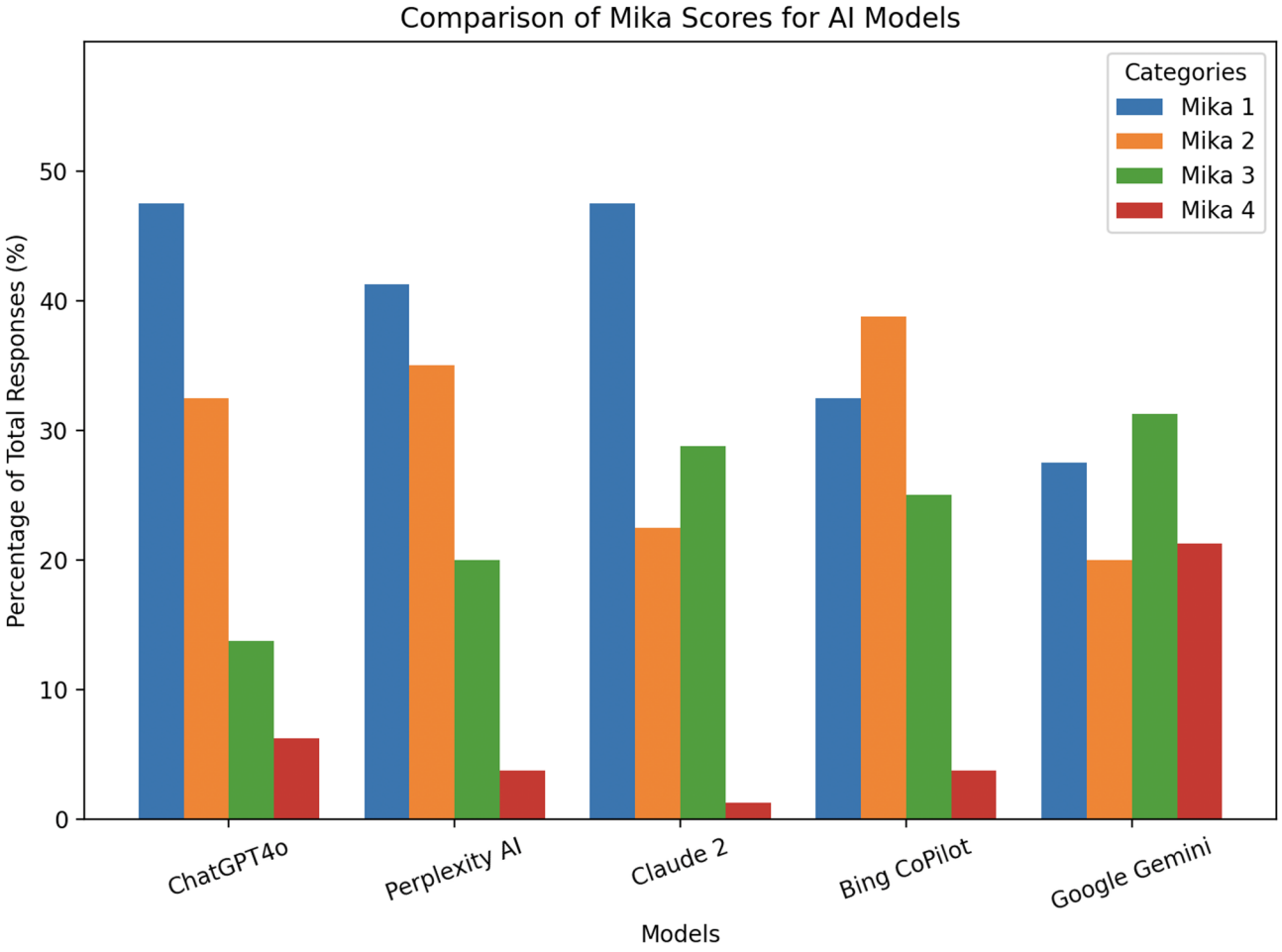

Fig. 1 Number of responses per large language model (LLM) categorized by the Mika score. Accuracy scores were defined as (1) "excellent response not requiring clarification"; (2) "satisfactory requiring minimal clarification"; (3) "satisfactory requiring moderate clarification"; or (4) "unsatisfactory requiring substantial clarification"

Full answers from each LLM are described in Supplementary Digital Material Tables 1, 2, 3, 4 and 5.

Inter-rater agreement

Responses to four LLMs had moderate agreement, including ChatGPT (AC2 = 0.435; 95%CI 0.421, 0.44), Perplexity AI (AC2 = 0.450; 95%CI 0.400, 0.500), Bing CoPilot (AC2 = 0.435; 95%CI 0.395, 0.476), and Claude2 (AC2 = 0.457; 95%CI 0.434, 0.481). There was no agreement for Google Gemini (AC2 = −0.039; 95%CI −0.076, −0.002).

Discussion

The primary finding of this study was that ChatGPT4o, Perplexity AI, Bing CoPilot, Claude2 generally provided "satisfactory responses requiring minimal clarification" to standardized questions relating to patellar instability and MPFLR. Google Gemini underperformed relative to these four LLMs, generally generating "satisfactory responses requiring moderate clarification", with a large portion of responses considered to be "unsatisfactory requiring substantial clarification". Responses from all LLMs apart from Google Gemini had moderate

agreement amongst reviewers, indicating heterogeneity in the perceptions of responses from different chatbots.

Over the past few years, there have been several different LLMs that have been created by various companies, each improving on a yearly basis in performance capabilities [22]. Despite this, research focused on assessing the accuracy of LLMs in answering medical questions consistently only evaluates ChatGPT [22]. This is likely due ChatGPT being the most mainstream version of all the LLMs [22]. This study suggests that for the purposes of answering questions related to patellar instability, Google Gemini underperforms compared to ChatGPT4o, Perplexity AI, Bing CoPilot, and Claude2, however findings should be interpreted with caution due to poor inter-rater reliability. There have been a variety of studies in various other disciplines such as ophthalmology and emergency medicine which have demonstrated that Gemini was outperformed by ChatGPT-4 [23–27]. However other studies such as that by Quinn et al. have demonstrated superiority with Gemini with regards to clarity of answers pertaining to ACLR [28]. Overall, as AI models are fundamentally free of judgment, patients may feel comfortable asking questions that they may perceive to be "unintelligent", facilitating clinic visits [29]. Therefore, ChatGPT4o, Perplexity AI, Bing CoPilot, and Claude2 are satisfactory adjuncts to surgeons in providing information regarding patellar instability and MPFLR, and are encouraged to be used by patients with this condition.

In addition to accuracy, inter-rater agreement is also an important metric when assessing the different models. Agreement between reviewers generally was over 0.400 using the Gwet AC2 coefficient in this study. In comparison, a similar study investigating the accuracy of ChatGPT in answering questions related to ACLR found an intra-class coefficient of 0.802 [8]. Another analysis found more similar results to the present study, inter-rater reliability values ranging from 0.28 to 0.43 for assessing the responses of ChatGPT for reliability, relevance, and accuracy in answering questions relating to total knee arthroplasty [30]. It is inevitable that reliability values will vary from study to study, even if they are investigating the same topic, given the different reviewers that assess the responses. It is unclear why Google Gemini had a low inter-rater reliability score, but it may be due to three reviewers scoring most of the responses from this chatbot as either a Mika 1 or 2. When these two are removed from analysis for Google Gemini, the Gwet AC2 coefficient increases to 0.509 (95%CI 0.420–0.598).

However, patellar instability and MPFLR is still a field that is rapidly evolving [13]. The Delphi Consensus study that the questions for this study were derived from concluded that more high-level evidence is necessary for development of universal standards [13]. While the ten prompts from this study were created from questions that had a strong consensus [13], there is still ongoing discourse about these aspects of patellar instability. For example, regarding the indications for operative versus nonoperative management, one recent systematic review reported a significant benefit in lowering redislocation rates with MPFLR for acute first-time patellar dislocation in the pediatric population (3.1% vs. 39.4%) [10]. Another systematic review reported a lower dislocation rate in patients with MPFLR compared to nonoperative management (7% vs. 30%) [11]. Other examples of controversy in this field include the use of isolated MPFLR for all patients (e.g. including patients with elevated tibial tubercle to trochlear groove (TT-TG)) distances [31], the use of tibial tubercle to posterior cruciate ligament (TT-PCL) distances versus TT-TG [32], measurement modality of TT-TG (e.g. MRI versus CT) [33], and an updated classification for trochlear dysplasia [34]. While the reasons for disagreement between quality of responses are multifactorial, the lack of definitive answers for various aspects of patellar instability in light of a lack of high quality evidence is likely a major contributor.

Several studies in general have described that LLMs provide satisfactory responses to questions related to orthopedics [8, 9, 35]. It can be concluded that for most orthopedic conditions, these LLMs are useful adjuncts to physicians and surgeons, and that patients should use them to gain information. As these softwares continue to develop, the potential for its usage will exponentially grow. The latest version of ChatGPT from OpenAI is now able to handle text, speech, and visual inputs, while also generating text, audio, and visual outputs. Using these new features to facilitate day-to-day workflow, analyze outcomes, streamline research, and enhance patient satisfaction can be the goal of future research in the AI space. In the context of PFI, future research should investigate the development of tailored chatbots trained on specific rehabilitation and postoperative protocols to provide reliable resources for patients.

This study is amongst few to assess the accuracy of five different LLMs within orthopedic research and patellar instability. Furthermore, all responses were assessed by a high number of practicing orthopedic surgeons, who all completed fellowship training within sports medicine. The questions chosen to prompt the LLMs were all from a previous Delphi Consensus study, all with strong consensus. Finally, all LLMs included in this study were free access, increasing generalizability. However, this study does have few limitations. First, surgeons were not blinded when reviewing responses from different LLMs, which may introduce bias. Second, despite moderate agreement as per Gwet's AC2 coefficient for most LLLMs, having humans assess the accuracy of responses introduces a level of bias within grading. Third, LLMs are updated frequently, and the quality of responses are likely

to improve with each subsequent update. Finally, this study only evaluated 10 questions related to PFI, limiting the ability to assess a broader and more diverse range of issues across this field. Capturing this phenomenon is not possible with a cross-sectional design.

Conclusion

Current free access LLMs (ChatGPT4o, Perplexity AI, Bing CoPilot, and Claude2) predominantly provide satisfactory responses requiring minimal clarification to standardized questions relating to patellar instability and MPFLR. Google Gemini statistically underperformed in accuracy relative to the other four LLMs, with most answers requiring moderate clarification. Furthermore, inter-rater agreement was moderate for all LLMs apart from Google Gemini. These findings advocate for the utility of existing LLMs in serving as an adjunct to physicians and surgeons in providing patients information pertaining to patellar instability and MPFLR.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12891-025-09227-1.

Supplementary Material 1

Acknowledgements
Not applicable.

Approval committee
Not applicable, there was no review board or approval committee involved as there were no human participants in this study.

Clinical trial number
Not applicable, the study is not a clinical trial.

Authors' contributions
P.V conceived of the study idea in conjunction with P.G, D.L, D.C.P.V. and D.S.E was responsible for the statistical analyses. Writing was done by P.V, D.C.K.N, M.T, B.M, L.H, R.L, B.L, A.M, S.T, P.S all contributed to reviewing the responses. D.D and K.S were supervisors and PIs. P.V, P.G, D.L, D.D, K.S, D.C reviewed the manuscript.

Funding
Open access funding provided by University of Gothenburg. Not applicable, there was no funding was provided for the development of this manuscript.

Data availability
Data may be made available upon reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable, there were no human patient participants in this research study. All authors have agreed to the publication of this manuscript.
Not applicable, there was no ethics board involved as there are no human patient participants in this research study. The only individuals involved in the manuscript were the authors.

Consent for publication
All authors have agreed to provide consent for publication

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Surgery, Division of Orthopaedic Surgery, McMaster University, 1200 Main St West, Hamilton, ON L8N 1H4, Canada
[2]Department of Orthopaedic Surgery, Kobe University Graduate School of Medicine, Kobe, Hyogo, Japan
[3]Department of Surgery, Division of Orthopaedic Surgery, University of Ottawa, Ottawa, ON, Canada
[4]Department of Surgery, Section of Orthopaedic Surgery, University of Calgary, Calgary, AB, Canada
[5]Department of Surgery, Section of Orthopaedic Surgery, University of Manitoba, Winnipeg, MB, Canada
[6]Department of Orthopaedic Surgery, University of Pittsburgh Medical Center, Pittsburgh, PA, USA
[7]Department of Orthopaedics and Trauma, Shalby Hospital Jaipur, Jaipur, India
[8]The Orthopaedic Specialty Clinic, Pune, Maharashtra, India
[9]Pihlajalinna Hospital, Tampere, Finland
[10]Department of Orthopaedics, Sahlgrenska University Hospital, Mölndal, Sweden
[11]Department of Orthopaedics, Institute of Clinical Sciences, the Sahlgrenska Academy, The University of Gothenburg, Gothenburg, Sweden

Received: 17 March 2025 / Accepted: 19 September 2025
Published online: 03 November 2025

References

1. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare. 2023;11:887. https://doi.org/10.3390/healthcare11060887.
2. Aljamaan F, Temsah M-H, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. JMIR Med Inform. 2024;12:e54345. https://doi.org/10.2196/54345.
3. Altıntaş E, Ozkent MS, Gül M, Batur AF, Kaynar M, Kılıç Ö, et al. Comparative analysis of artificial intelligence chatbot recommendations for urolithiasis management: a study of EAU guideline compliance. The French Journal of Urology. 2024;34:102666. https://doi.org/10.1016/j.fjurol.2024.102666.
4. Martin RK, Wastvedt S, Pareek A, Persson A, Visnes H, Fenstad AM, et al. Predicting anterior cruciate ligament reconstruction revision: a machine learning analysis utilizing the Norwegian knee ligament register. J Bone Joint Surg Am. 2022;104:145–53. https://doi.org/10.2106/JBJS.21.00113.
5. Pareek A, Ro DH, Karlsson J, Martin RK. Machine learning/artificial intelligence in sports medicine: state of the art and future directions. Journal of ISAKOS. 2024. https://doi.org/10.1016/j.jisako.2024.01.013.
6. Sniderman J, Stark RB, Schwartz CE, Imam H, Finkelstein JA, Nousiainen MT. Patient factors that matter in predicting hip arthroplasty outcomes: a machine-learning approach. J Arthroplasty. 2021;36:2024–32. https://doi.org/10.1016/j.arth.2020.12.038.
7. AlShehri Y, McConkey M, Lodhia P. ChatGPT has educational potential: assessing ChatGPT responses to common patient hip arthroscopy questions. Arthroscopy. 2024;S0749-8063(24):00452–3. https://doi.org/10.1016/j.arthro.2024.06.017.
8. Johns WL, Kellish A, Farronato D, Ciccotti MG, Hammoud S. ChatGPT can offer satisfactory responses to common patient questions regarding elbow ulnar collateral ligament reconstruction. Arthroscopy, Sports Medicine, and Rehabilitation. 2024;6:100893. https://doi.org/10.1016/j.asmr.2024.100893.
9. Li LT, Sinkler MA, Adelstein JM, Voos JE, Calcei JG. ChatGPT responses to common questions about anterior cruciate ligament reconstruction are frequently satisfactory. Arthroscopy. 2024;40:2058–66. https://doi.org/10.1016/j.arthro.2023.12.009.
10. Blackman B, Dworsky-Fried J, Cohen D, Slawaska-Eng D, Gyemi L, Simunovic N, et al. Surgical management of first-time patellar dislocations in paediatric patients may lower rates of redislocation compared to conservative management: a systematic review and meta-analysis. Knee Surg Sports Traumatol Arthrosc. 2024. https://doi.org/10.1002/ksa.12524.

11. Cohen D, Le N, Zakharia A, Blackman B, de Sa D. MPFL reconstruction results in lower redislocation rates and higher functional outcomes than rehabilitation: a systematic review and meta-analysis. Knee Surg Sports Traumatol Arthrosc. 2022;30:3784–95. https://doi.org/10.1007/s00167-022-07003-5.
12. Ponkilainen V, Kuitunen I, Liukkonen R, Vaajala M, Reito A, Uimonen M. The incidence of musculoskeletal injuries: a systematic review and meta-analysis. Bone Joint Res. 2022;11:814–25. https://doi.org/10.1302/2046-3758.1111.BJR-2022-0181.R1.
13. Hurley ET, Hughes AJ, Savage-Elliott I, Dejour D, Campbell KA, Mulcahey MK, et al. A modified Delphi consensus statement on patellar instability: part I. Bone Jt J. 2023;105–B:1259–64. https://doi.org/10.1302/0301-620X.105B12.BJJ-2023-0109.R1.
14. Hurley ET, Sherman SL, Chahla J, Gursoy S, Alaia MJ, Tanaka MJ, et al. A modified Delphi consensus statement on patellar instability: part II. Bone Jt J. 2023;105–B:1265–70. https://doi.org/10.1302/0301-620X.105B12.BJJ-2023-0110.R1.
15. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing chatgpt responses to common patient questions regarding total hip arthroplasty. J Bone Joint Surg Am. 2023;105:1519–26. https://doi.org/10.2106/JBJS.23.00209.
16. Mika AP, Mulvey HE, Engstrom SM, Polkowski GG, Martin JR, Wilson JM. Can chatgpt answer patient questions regarding total knee arthroplasty? J Knee Surg. 2024;37:664–73. https://doi.org/10.1055/s-0044-1782233.
17. Johns WL, Martinazzi BJ, Miltenberg B, Nam HH, Hammoud S. ChatGPT Provides Unsatisfactory Responses to Frequently Asked Questions Regarding Anterior Cruciate Ligament Reconstruction. Arthroscopy. 2024;40:2067–2079.e1. https://doi.org/10.1016/j.arthro.2024.01.017.
18. Wrenn SP, Mika AP, Ponce RB, Mitchell PM. Evaluating chatgpt's ability to answer common patient questions regarding hip fracture. J Am Acad Orthop Surg. 2024;32:656–9. https://doi.org/10.5435/JAAOS-D-23-00877.
19. Jeyaraman MM, Rabbani R, Copstein L, Robson RC, Al-Yousif N, Pollock M, et al. Methodologically rigorous risk of bias tools for nonrandomized studies had low reliability and high evaluator burden. J Clin Epidemiol. 2020;128:140–7. https://doi.org/10.1016/j.jclinepi.2020.09.033.
20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
21. Lorenz RC, Matthias K, Pieper D, Wegewitz U, Morche J, Nocon M, et al. A psychometric study found AMSTAR 2 to be a valid and moderately reliable appraisal tool. J Clin Epidemiol. 2019;114:133–40. https://doi.org/10.1016/j.jclinepi.2019.05.028.
22. Nwachukwu BU, Varady NH, Allen AA, Dines JS, Altchek DW, Williams RJ et al. Currently available large Language models do not provide musculoskeletal treatment recommendations that are concordant with Evidence-Based clinical practice guidelines. Arthroscopy. 2024;S0749-8063(24):00575-9. https://doi.org/10.1016/j.arthro.2024.07.040.
23. Abdul Sami M, Abdul Samad M, Parekh K, Suthar PP. Comparative accuracy of ChatGPT 4.0 and Google Gemini in answering pediatric radiology text-based questions. Cureus. 2024;16:e70897. https://doi.org/10.7759/cureus.70897.
24. Fujimoto M, Kuroda H, Katayama T, Yamaguchi A, Katagiri N, Kagawa K, et al. Evaluating large language models in dental anesthesiology: a comparative analysis of ChatGPT-4, Claude 3 Opus, and gemini 1.0 on the Japanese dental society of anesthesiology board certification exam. Cureus. 2024;16:e70302. https://doi.org/10.7759/cureus.70302.
25. Lee TJ, Campbell DJ, Patel S, Hossain A, Radfar N, Siddiqui E, et al. Unlocking health literacy: the ultimate guide to hypertension education from ChatGPT versus Google gemini. Cureus. 2024;16:e59898. https://doi.org/10.7759/cureus.59898.
26. Masalkhi M, Ong J, Waisberg E, Lee AG. Google DeepMind's Gemini AI versus ChatGPT: a comparative analysis in ophthalmology. Eye. 2024;38:1412–7. https://doi.org/10.1038/s41433-024-02958-w.
27. Tong L, Zhang C, Liu R, Yang J, Sun Z. Comparative performance analysis of large language models: ChatGPT-3.5, ChatGPT-4 and Google Gemini in glucocorticoid-induced osteoporosis. J Orthop Surg Res. 2024;19:574. https://doi.org/10.1186/s13018-024-04996-2.
28. Quinn M, Milner JD, Schmitt P, Morrissey P, Lemme N, Marcaccio S et al. Artificial intelligence large Language models address anterior cruciate ligament reconstruction: superior clarity and completeness by gemini compared with ChatGPT-4 in response to American academy of orthopaedic surgeons clinical practice guidelines. Arthroscopy. 2024;S0749-8063(24):00736-9. https://doi.org/10.1016/j.arthro.2024.09.020.
29. Kunze KN, Jang SJ, Fullerton MA, Vigdorchik JM, Haddad FS. What's all the chatter about? Bone Jt J. 2023;105–B:587–9. https://doi.org/10.1302/0301-620X.105B6.BJJ-2023-0156.
30. Magruder ML, Rodriguez AN, Wong JCJ, Erez O, Piuzzi NS, Scuderi GR, et al. Assessing ability for ChatGPT to answer total knee Arthroplasty-related questions. J Arthroplasty. 2024;39:2022–7. https://doi.org/10.1016/j.arth.2024.02.023.
31. Vivekanantha P, Kahlon H, Cohen D, de Sa D. Isolated medial patellofemoral ligament reconstruction results in similar postoperative outcomes as medial patellofemoral ligament reconstruction and tibial-tubercle osteotomy: a systematic review and meta-analysis. Knee Surg Sports Traumatol Arthrosc. 2023;31:2433–45. https://doi.org/10.1007/s00167-022-07186-x.
32. Vivekanantha P, Kahlon H, Shahabinezhad A, Cohen D, Nagai K, Hoshino Y, et al. Tibial tubercle to trochlear groove distance versus tibial tubercle to posterior cruciate ligament distance for predicting patellar instability: a systematic review. Knee Surg Sports Traumatol Arthrosc. 2023;31:3243–58. https://doi.org/10.1007/s00167-023-07358-3.
33. Camp CL, Stuart MJ, Krych AJ, Levy BA, Bond JR, Collins MS, et al. CT and MRI measurements of tibial tubercle-trochlear groove distances are not equivalent in patients with patellar instability. Am J Sports Med. 2013;41(8):1835–40. https://doi.org/10.1177/0363546513484895.
34. Dejour DH, de Sanctis EG, Müller JH, Deroche E, Pineda T, Guarino A, et al. Adapting the dejour classification of trochlear dysplasia from qualitative radiograph- and CT-based assessments to quantitative MRI-based measurements. Knee Surg Sports Traumatol Arthrosc. 2025;33(8):2833–46. https://doi.org/10.1002/ksa.12539.
35. Dagher T, Dwyer EP, Baker HP, Kalidoss S, Strelzow JA. Dr. AI will see you now: how do ChatGPT-4 treatment recommendations align with orthopaedic clinical practice guidelines? Clin Orthop. 2024. https://doi.org/10.1097/CORR.0000000000003234.

Publisher's Note