





Article

AI in Hand Surgery: Assessing Large Language Models in the Classification and Management of Hand Injuries

Sophia M. Pressman ¹, Sahar Borna ¹, Cesar A. Gomez-Cabello ¹, Syed Ali Haider ¹
and Antonio Jorge Forte ^{1,2,*}

¹ Division of Plastic Surgery, Mayo Clinic, Jacksonville, FL 32224, USA; sophiampressman1@gmail.com (S.M.P.); saharborna2021@gmail.com (S.B.); abrahamgomezcabello@gmail.com (C.A.G.-C.); dr.s.alihaider@gmail.com (S.A.H.)

² Center for Digital Health, Mayo Clinic, Rochester, MN 55905, USA

* Correspondence: ajvforte@yahoo.com.br

Abstract: Background: OpenAI's ChatGPT (San Francisco, CA, USA) and Google's Gemini (Mountain View, CA, USA) are two large language models that show promise in improving and expediting medical decision making in hand surgery. Evaluating the applications of these models within the field of hand surgery is warranted. This study aims to evaluate ChatGPT-4 and Gemini in classifying hand injuries and recommending treatment. **Methods:** Gemini and ChatGPT were given 68 fictionalized clinical vignettes of hand injuries twice. The models were asked to use a specific classification system and recommend surgical or nonsurgical treatment. Classifications were scored based on correctness. Results were analyzed using descriptive statistics, a paired two-tailed *t*-test, and sensitivity testing. **Results:** Gemini, correctly classifying 70.6% hand injuries, demonstrated superior classification ability over ChatGPT (mean score 1.46 vs. 0.87, *p*-value < 0.001). For management, ChatGPT demonstrated higher sensitivity in recommending surgical intervention compared to Gemini (98.0% vs. 88.8%), but lower specificity (68.4% vs. 94.7%). When compared to ChatGPT, Gemini demonstrated greater response replicability. **Conclusions:** Large language models like ChatGPT and Gemini show promise in assisting medical decision making, particularly in hand surgery, with Gemini generally outperforming ChatGPT. These findings emphasize the importance of considering the strengths and limitations of different models when integrating them into clinical practice.

Keywords: artificial intelligence (AI); ChatGPT; Gemini; deep learning; machine learning; hand surgery; hand trauma; management



Citation: Pressman, S.M.; Borna, S.; Gomez-Cabello, C.A.; Haider, S.A.; Forte, A.J. AI in Hand Surgery: Assessing Large Language Models in the Classification and Management of Hand Injuries. *J. Clin. Med.* **2024**, *13*, 2832. <https://doi.org/10.3390/jcm13102832>

Academic Editor: Steffen U. Eisenhardt

Received: 11 April 2024

Revised: 29 April 2024

Accepted: 9 May 2024

Published: 11 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The integration of artificial intelligence (AI) into daily medical practice is an evitable reality [1,2]. With an expanding repertoire of applications and tools, AI is infiltrating the healthcare landscape, seeping into every specialty and healthcare domain [3,4]. This progression is marked by the emergence of large language models (LLMs). Built upon neural network architectures [5], LLMs are AI systems which are designed to employ deep learning and natural language processing [6]. These models undergo extensive training on vast datasets until they are able to comprehend and generate human-like text with significant predictability [6], even without employing retrieval-augmented generation (RAG) approaches. Two publicly available LLMs that have garnered recent interest are Google's Gemini [7] (built upon its predecessor Bard) and OpenAI's ChatGPT [8]. Although not designed specifically for medical purposes, these LLMs have demonstrated potential in medical decision making [9]. For example, ChatGPT achieved passing scores for both the United States Medical Licensing Examination [10] and American Society for Surgery of the Hand self-assessment exam [11], demonstrating a basic level of competence in medical

reasoning. With the potential to enhance diagnostic accuracy, informed decision making, and patient outcomes, LLMs like ChatGPT and Gemini show promise in augmenting healthcare delivery [9,12].

The field of hand surgery stands to benefit greatly from the support of AI tools like LLMs [1,13]. Achieving favorable outcomes after hand trauma is paramount, given the critical role that hands play in necessary tasks for daily functioning and independence [14–17]. Effective management relies heavily on correctly identifying the nature and severity of injuries, as this information guides decisions regarding surgical intervention, rehabilitation, and ongoing care [18–21]. Inaccurate diagnosis and classification can lead to delays in treatment, inappropriate interventions, and potentially compromised functional recovery for patients [22,23]. With their advanced algorithms, LLMs can analyze complex patterns [6] and thus have the potential to provide rapid and precise injury classifications, enhancing and expediting the management process.

While the integration of advanced LLMs into healthcare systems is on the horizon, access may be limited to those within those institutions. Healthcare professionals in underfunded facilities, medical students, and patients are expected to continue using publicly available LLMs like ChatGPT and Gemini for medical queries. The continuous assessment of the value and utility of these LLMs is essential. Early investigations into LLM applications in hand surgery have shown promising results. Leypold et al. [24] found that ChatGPT adeptly managed complex hand and arm surgery scenarios, showcasing the potential of LLMs like ChatGPT to improve patient care and surgical outcomes. Crook et al. [25] reported ChatGPT's proficiency in addressing common patient inquiries regarding common hand surgeries, noting generally high-quality responses. In a similar study, Seth et al. [26] found that ChatGPT was suitable for nonmedical individuals, but struggled with accurate and complete references. Additionally, Al Rawi et al. [27] observed that, while ChatGPT's responses were correct and useful in most cases, only 57% were deemed complete by hand surgeon reviewers.

Although these studies show promise for ChatGPT, there is limited research exploring Gemini's potential in hand surgery and few studies comparing LLMs. Furthermore, additional research to explore the full extent of LLM capabilities in the context of hand injuries and hand surgery, particularly in specialized tasks like injury classification, is warranted. The objective of this study is to assess the ability of ChatGPT and Gemini to accurately classify hand injuries via the use of 12 specific classification systems. Furthermore, this study aims to determine the ability of these models to accurately recommend surgical or nonsurgical management for these hand injuries. In doing so, we seek to evaluate the capabilities of publicly available LLMs without the use of RAG approaches. Through this investigation, this study endeavors to advance the ongoing discourse surrounding the applications and limitations of LLMs in hand surgery.

2. Materials and Methods

2.1. Study Design

Sixty-eight unique prompts were developed to test each LLM's ability to classify hand injuries. Prompts cover 12 different classification systems [20,21,28–39] covering various hand injuries. The inclusion criteria prioritized classification systems with well-established significance, ensuring relevance and familiarity among healthcare providers specializing in hand surgery. To focus on clinically relevant systems, classification systems lacking direct treatment correlations were excluded.

Each prompt was prefaced with, "I am a plastic and reconstructive surgeon who specializes in hand surgery. You are my colleague and I am discussing a case with you." Each prompt included a fictionalized vignette and a specific hand injury diagnosis. Additionally, within each prompt was a request to classify the injury using a specific classification system and determine if this injury warrants surgical or nonsurgical (conservative) management. The deliberate inclusion of the specific diagnosis and classification system of interest was incorporated to focus the evaluation of the LLMs' classification abilities, rather than their

diagnostic skills, ensuring a uniform methodology and reducing the likelihood of LLMs using varying classification systems. Examples of prompts with LLM responses are depicted in Figures 1 and 2. Each prompt was provided to ChatGPT-4 (OpenAI, San Francisco, CA, USA) and Gemini (Google, Mountain View, CA, USA) twice to ensure consistency and replicability, resulting in a total of 136 prompts. Additionally, each prompt was entered individually in a separate conversation to minimize the possibility of one answer affecting another. All prompts were provided on 10 March 2024, using the Google Chrome internet browser.

ChatGPT-4

You
I am a plastic and reconstructive surgeon who specializes in hand surgery. You are my colleague and I am discussing a case with you.

A 68-year-old female presents for evaluation of hand pain immediately after a forward fall down the stairs on her outstretched hand. Signs and symptoms are most consistent with a stable scaphoid fracture. Imaging reveals a nondisplaced waist fracture.

Based on this patient's presentation, answer the following 2 questions:
Question 1. Using the Herbert and Fisher Classification of scaphoid fractures, how would you classify this injury?
Question 2. For this injury do you recommend (A) non-surgical (conservative) management OR (B) surgical management/intervention?

For both answers, you must be as specific and decisive as possible. Do not provide multiple options (provide only 1 answer for each question).

ChatGPT
Given the information about the 68-year-old female patient who suffered a forward fall resulting in a nondisplaced waist fracture of the scaphoid, here are my specific recommendations based on the Herbert and Fisher Classification, as well as current management guidelines for such injuries:

Question 1: Classification of the Injury
Using the Herbert and Fisher Classification, this injury would be classified as a Type B2 fracture. This classification system is specifically designed for scaphoid fractures and places a particular emphasis on the fracture's location, displacement, and potential for instability. A Type B2 fracture refers to a stable, nondisplaced fracture at the waist of the scaphoid. This type of fracture is known for having a good prognosis with appropriate management due to its stable nature and the relatively good blood supply in this area, which is conducive to healing.

Question 2: Recommended Management Strategy
For a stable, nondisplaced waist fracture of the scaphoid in an elderly patient, I recommend (A) non-surgical (conservative) management. This decision is based on several factors:

- The nature of the fracture being stable and nondisplaced suggests a high potential for healing with conservative measures.
- Conservative management typically involves immobilization of the wrist in a cast or splint for a period, usually between 6 to 12 weeks, allowing the bone to heal naturally.
- Considering the patient's age, non-surgical management minimizes the risks associated with surgery and anesthesia, which can be particularly beneficial unless there are specific indications for surgical intervention.

It's important to monitor the patient closely during the conservative management phase with regular imaging to ensure proper healing of the fracture and to reassess the management plan if healing does not progress as expected. Early and consistent follow-ups are crucial to detect any potential complications or failure to heal, which might then necessitate reconsideration of surgical options.

Gemini

You
I am a plastic and reconstructive surgeon who specializes in hand surgery. You are my colleague and I am discussing a case with you.

A 68-year-old female presents for evaluation of hand pain immediately after a forward fall down the stairs on her outstretched hand. Signs and symptoms are most consistent with a stable scaphoid fracture. Imaging reveals a nondisplaced waist fracture.

Based on this patient's presentation, answer the following 2 questions:
Question 1. Using the Herbert and Fisher Classification of scaphoid fractures, how would you classify this injury?
Question 2. For this injury do you recommend (A) non-surgical (conservative) management OR (B) surgical management/intervention?

For both answers, you must be as specific and decisive as possible. Do not provide multiple options (provide only 1 answer for each question).

Show drafts ▾ 🔊

◆ Answer 1: Based on the Herbert and Fisher Classification of scaphoid fractures, this injury would be classified as **A2: acute stable undisplaced waist fracture**.

Answer 2: For this injury, I recommend (A) **non-surgical (conservative) management**.

Here's the reasoning:

- The patient has a stable scaphoid fracture, evident by the nondisplaced waist fracture and the classification of A2 in the Herbert and Fisher system.
- Stable fractures, like A2, typically heal well with conservative management due to good blood supply in the area.

Conservative management for this case would likely involve immobilization with a cast for a period of time, followed by physical therapy to regain strength and motion in the wrist.

👍 🗨️ 🌐 ⏪ ⏩ 🌍 ⋮

Figure 1. An example of a prompt given to ChatGPT-4 (left) and Gemini (right) with the corresponding responses below. This prompt asked the models to classify a scaphoid fracture using Herbert and Fisher’s classification system.

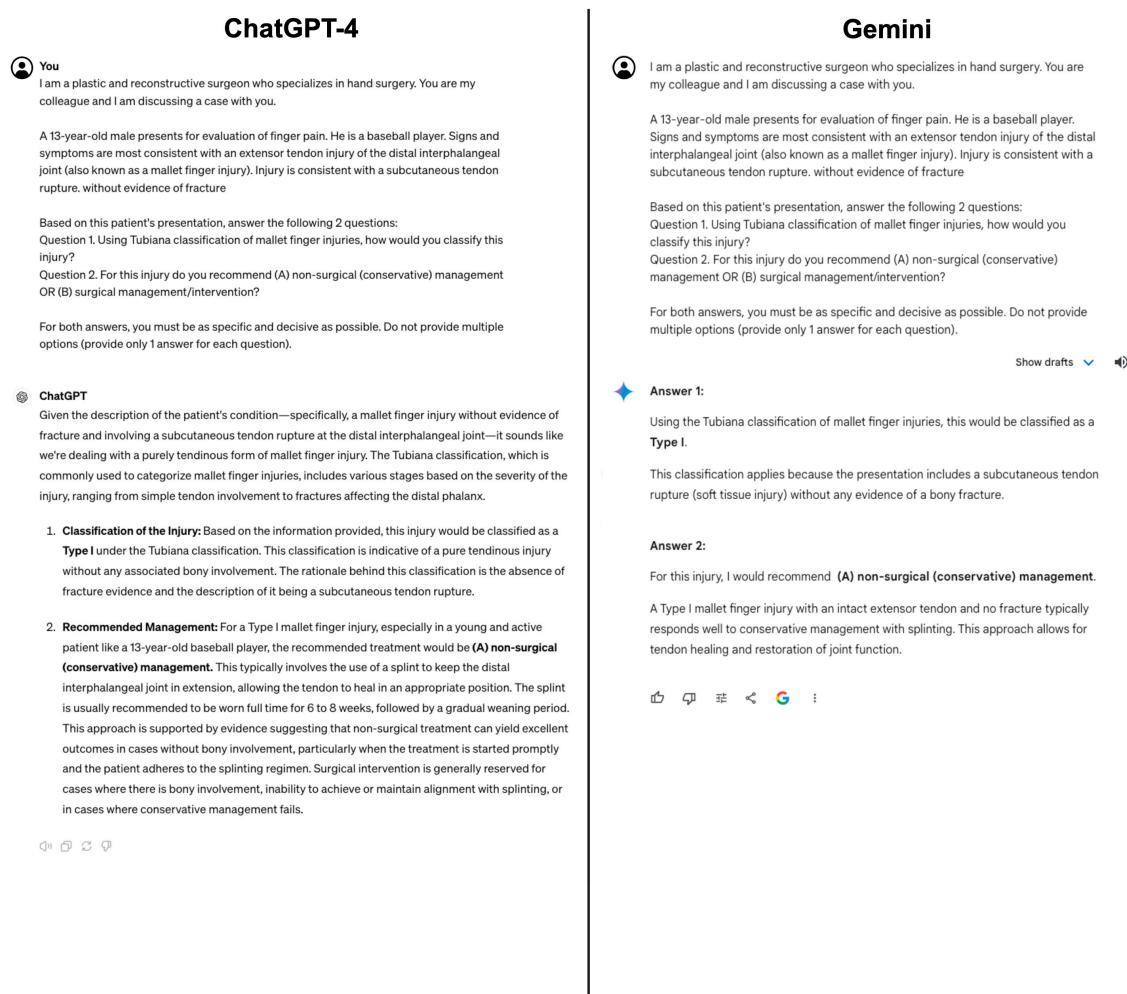


Figure 2. An example of a prompt given to ChatGPT-4 (left) and Gemini (right) with the corresponding responses. This prompt asked the models to classify a mallet finger injury using Tubiana’s classification system.

For the assessment of classification abilities, the LLMs were given points based on answer correctness. Completely correct classifications were awarded two points, partially correct classifications were awarded one point, and incorrect classifications were awarded zero points. Instances of partial correctness typically arose from either indecisiveness on the part of the LLM or from including subclassifications within the classification system. For example, in systems like the Gustilo–Anderson classification [20,33] for open fractures, which have subclassifications (e.g., Type IIIA, IIIB, and IIIC), partial correctness could result from selecting the wrong subclassification (e.g., selecting Type IIIB when the correct answer was Type IIIA) or from failing to specify the subclassification altogether (e.g., answering Type III when the correct answer was Type IIIA). Furthermore, if the LLM struggled to choose between two options, one of which was correct, it would still receive one point for partial correctness.

In our investigation into the LLM’s ability to differentiate between surgical and nonsurgical management options for hand injuries, we relied on clinically pertinent classification systems as our gold standard. These classification systems were chosen for their comprehensive treatment recommendations which were associated with each categorized injury. As such, the recommendations provided within the classification systems served as the definitive ‘correct’ answers for our study. Furthermore, to ensure the accuracy of our assessments, all responses were verified by a board-certified hand surgeon. By adhering to these established guidelines, we aimed to maintain consistency and objectivity in our

evaluation process. This approach not only provided a clear framework for comparing LLM performance, but also ensured the clinical relevance and validity of our findings.

2.2. Data Collection and Analysis

Prompts and corresponding LLM responses were collected in Microsoft Excel (Redmond, WA, USA). Each response was graded based on classification correctness. These values were analyzed using descriptive statistics including the mean, standard deviation (SD), and range. Comparisons between ChatGPT and Gemini classification abilities were evaluated using paired, two-tailed *t*-tests. This test controls for variability across different clinical vignettes and allows for the detection of significant differences, without assuming directionality. Subgroup analyses were conducted to compare scores for each classification system. An alpha level of 0.05 was used to determine statistical significance.

In addition to classification accuracy, the evaluation of the LLMs' capacity to recommend surgical versus nonsurgical management involved the calculation of sensitivity (also known as recall), specificity, positive predictive value (PPV; also known as precision), accuracy, and the F1 score. Sensitivity and specificity offer insights into the models' ability to accurately identify true surgical and nonsurgical cases, respectively, while PPV and accuracy shed light on the precision and correctness of the models' recommendations. Furthermore, the F1 score serves as a synthesized measure, capturing the balance between precision and recall, and offering a more nuanced understanding of the models' overall performance. This multifaceted evaluation approach ensures that the strengths and weaknesses of the LLMs' recommendations are thoroughly explored, providing valuable insights for healthcare providers and researchers alike in navigating the complexities of decision making in hand surgery.

3. Results

3.1. Classification Results

In the classification of hand injuries, Gemini exhibited superior performance over ChatGPT, with an average score of 1.46 (SD 0.87), whereas ChatGPT yielded an average score of 0.67 (SD 0.87) (*p*-value < 0.001). Classification results are displayed in Table 1 and Figure 3. Gemini provided completely correct classifications for 96 (70.6%) hand injuries, but partially correct and incorrect for six (4.4%) and 34 (25.0%) injuries, respectively. ChatGPT provided correct, partially correct, and incorrect classifications for 36 (26.5%), 19 (14.0%), and 81 (59.6%) hand injuries, respectively.

ChatGPT exhibited its strongest performance in utilizing the Lichtman classification [38] for Kienböck disease (osteonecrosis of the lunate), correctly classifying eight cases (66.7%) with a mean score of 1.58 (SD 0.67). Additionally, its next best performance was in employing the Gustilo–Anderson classification [20,33] for open fractures, accurately categorizing seven cases (70%) with a mean score of 1.50 (SD 0.85). However, its performance was notably poorer when using Hintermann et al.'s classification [35] system for Gamekeeper's thumb, where all of its classifications were incorrect. Similarly, ChatGPT struggled with the classification of scaphoid fractures, where it inaccurately classified all cases according to the Mayo classification [28,29] system. Furthermore, in classifying volar plate avulsion injuries using the Eaton classification [30] system, ChatGPT demonstrated unacceptable performance, failing to correctly classify any cases. In contrast, Gemini displayed superior classification capabilities, providing accurate classifications for all volar plate avulsion injuries using the Eaton classification [30] system and all scaphoid fractures using the Herbert and Fisher Classification [34]. Despite this, Gemini's weakest performance was observed when classifying flexor tendon injuries using Kleinert and Verdan's Zone classification [36] system. Notably, ChatGPT narrowly outperformed Gemini only in the Gustilo–Anderson classification [20,33] of open fractures, Kleinert and Verdan's Zone classification [36] of flexor tendon injuries, and the Lichtman classification [38] for Kienböck disease (osteonecrosis of the lunate).

Table 1. Hand injury classification results from the LLMs.

| Classification System | LLM | Correct | Partially Correct | Incorrect | Mean Score | Standard Deviation |
|--|-----------|---------|-------------------|-----------|------------|--------------------|
| Eaton classification for volar plate avulsion injuries | ChatGPT-4 | 0 | 1 | 7 | 0.13 | 0.35 |
| | Gemini | 8 | 0 | 0 | 2.00 | 0.00 |
| Geissler arthroscopic classification for carpal instability | ChatGPT-4 | 3 | 0 | 5 | 0.75 | 1.04 |
| | Gemini | 4 | 0 | 4 | 1.00 | 1.07 |
| Green and O'Brien's classification of thumb metacarpal fractures | ChatGPT-4 | 1 | 1 | 10 | 0.25 | 0.62 |
| | Gemini | 11 | 1 | 0 | 1.92 | 0.29 |
| Gustilo-Anderson classification of open fractures | ChatGPT-4 | 7 | 1 | 2 | 1.50 | 0.85 |
| | Gemini | 5 | 3 | 2 | 1.30 | 0.82 |
| Herbert and Fisher Classification of scaphoid fractures | ChatGPT-4 | 4 | 6 | 10 | 0.70 | 0.80 |
| | Gemini | 20 | 0 | 0 | 2.00 | 0.00 |
| Hintermann et al.'s classification of ulnar collateral ligament (UCL) injury of the thumb | ChatGPT-4 | 0 | 0 | 12 | 0.00 | 0.00 |
| | Gemini | 10 | 0 | 2 | 1.67 | 0.78 |
| Kleinert and Verdan's Zone classification of flexor tendon injuries | ChatGPT-4 | 3 | 6 | 7 | 0.75 | 0.77 |
| | Gemini | 4 | 2 | 10 | 0.63 | 0.89 |
| Leddy and Packer classification of avulsion injury of the flexor digitorum profundus (FDP) | ChatGPT-4 | 4 | 0 | 8 | 0.67 | 0.98 |
| | Gemini | 6 | 0 | 6 | 1.00 | 1.04 |
| Lichtman classification of Kienböck disease (osteonecrosis the lunate) | ChatGPT-4 | 8 | 3 | 1 | 1.58 | 0.67 |
| | Gemini | 8 | 0 | 4 | 1.33 | 0.98 |
| Mayfield classification for carpal instability | ChatGPT-4 | 4 | 1 | 3 | 1.13 | 0.99 |
| | Gemini | 6 | 0 | 2 | 1.50 | 0.93 |
| Mayo Classification of scaphoid fractures | ChatGPT-4 | 0 | 0 | 10 | 0.00 | 0.00 |
| | Gemini | 8 | 0 | 2 | 1.60 | 0.84 |
| Tubiana classification for mallet finger | ChatGPT-4 | 2 | 0 | 6 | 0.50 | 0.93 |
| | Gemini | 6 | 0 | 2 | 1.50 | 0.93 |
| Total | ChatGPT-4 | 36 | 19 | 81 | 0.67 | 0.87 |
| | Gemini | 96 | 6 | 34 | 1.46 | 0.87 |

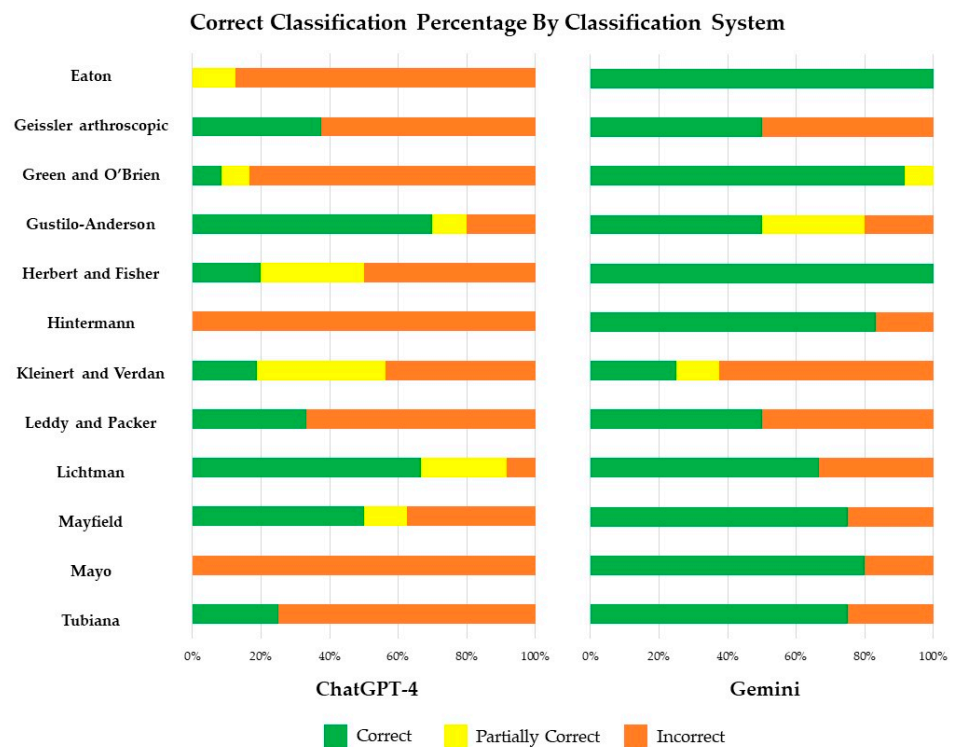


Figure 3. Percentage of correct classifications for each classification system for ChatGPT-4 (left) and Gemini (right).

3.2. Surgical Management Results from Sensitivity Testing

Based on the clinical vignettes and classification, 98 (72.1%) injuries warranted surgical intervention, and 38 (27.9%) justified nonsurgical management. ChatGPT recommended surgical intervention for 108 (79.4%) cases, as compared to Gemini, which recommended surgery for only 89 (65.4%). The results of the sensitivity testing are shown in Table 2. ChatGPT demonstrated higher sensitivity (recall) in recommending surgical intervention when compared to Gemini (98.0% vs. 88.8%). However, Gemini demonstrated a specificity of 94.7%, which was higher than ChatGPT’s specificity of 68.4%. Additionally, Gemini demonstrated a PPV (precision) of 97.8%, which was higher than ChatGPT’s PPV of 88.9%. Both models exhibited comparable F1 scores, with ChatGPT achieving an F1 score of 0.932 and Gemini achieving an F1 score of 0.930.

Table 2. Results of sensitivity testing.

| Value | ChatGPT | Gemini |
|---------------------------------|---------|--------|
| Sensitivity | 0.980 | 0.888 |
| Specificity | 0.684 | 0.947 |
| Positive Predictive Value (PPV) | 0.889 | 0.978 |
| Negative Predictive Value (NPV) | 0.929 | 0.766 |
| Positive Likelihood Ratio (LR+) | 3.102 | 16.867 |
| Negative Likelihood Ratio (LR−) | 0.030 | 0.118 |
| Accuracy | 0.897 | 0.904 |
| F1 score | 0.932 | 0.930 |

3.3. Replicability Results

To ensure consistency, each of the 68 prompts was presented twice. Gemini’s classification response differed in six instances (8.9%). In contrast, ChatGPT showed more variability, with its classification changing in 17 cases (25.0%), indicating a lower level of consistency and replicability. Among ChatGPT’s changes, 12 (70.5%) resulted in a more accurate classification (e.g., changing from incorrect to partially correct, changing from incorrect to correct, or changing from partially correct to correct). In terms of recommending surgical or nonsurgical management, ChatGPT modified its answer for six (8.9%) injuries, five of which were corrected in the subsequent response. However, Gemini’s response changed only once (1.5%), albeit to the incorrect management choice.

4. Discussion

With the correct classification of 70.6% of hand injuries, Gemini demonstrated superior performance to ChatGPT, which correctly classified just over a quarter of the injuries. ChatGPT’s poor performance, specifically when using the Eaton [30], Hintermann [35], and Mayo [28,29] classification systems, may suggest a lack of information relating to these classification systems in the dataset on which it was trained. The results of this study demonstrate Gemini’s superior performance over ChatGPT, which contrasts the few prior comparative studies involving Gemini, where ChatGPT was found to demonstrate greater accuracy [40,41]. Similarly, previous studies comparing ChatGPT to Gemini’s predecessor, Bard, have shown varied results, with some studies favoring ChatGPT [42–45] and others favoring Bard [46,47]. Without a definitive accuracy threshold, neither model is currently reliable enough as a classification tool to be used in clinical practice, but this situation is expected to change shortly. While neither ChatGPT nor Gemini were designed for medical use, the findings reveal promising potential in hand injury classification, which is just one of its potential applications in hand surgery (Figure 4). With the further expansion of datasets and the refinement of algorithms, these models are anticipated to reach the required level of accuracy for practical use.

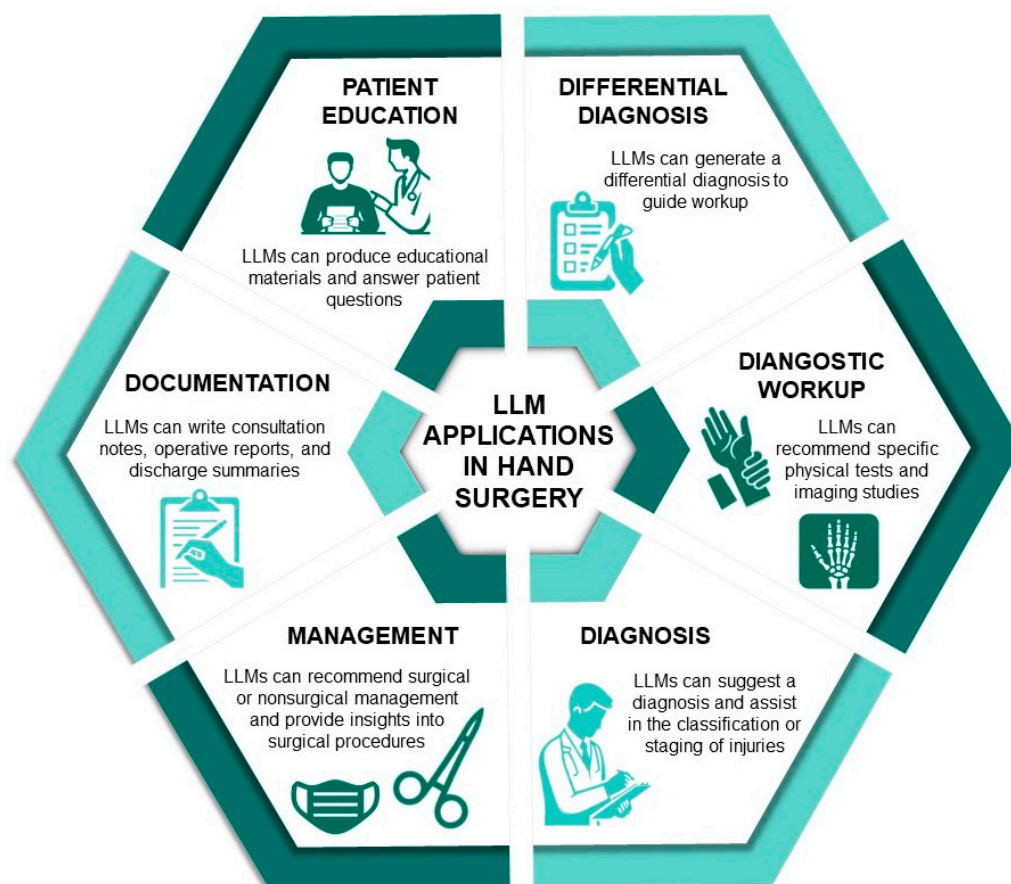


Figure 4. Applications of large language models (LLMs) in hand surgery.

The findings from sensitivity testing indicate that ChatGPT leans towards recommending surgical intervention more quickly, whereas Gemini takes a more cautious and conservative approach, showing hesitation in recommending surgical intervention. These differing behaviors carry significant implications, particularly in clinical decision-making scenarios. While ChatGPT's promptness may offer a sense of urgency, it also raises concerns regarding the potential for the over-recommendation of surgical procedures. On the other hand, Gemini's reluctance may contribute to a more conservative approach, minimizing the risk of unnecessary interventions, but possibly at the expense of timely action when surgical intervention is indeed warranted. Healthcare providers must carefully consider these nuances when incorporating AI decision-support tools into clinical practice, balancing the need for prompt action with the importance of exercising prudence and minimizing unnecessary interventions.

When compared to ChatGPT, Gemini also demonstrated greater consistency in its answers. This is not unsurprising, as previous studies [48–50] have reported concerns regarding ChatGPT's consistency. To be a reliable medical resource, an LLM must ensure that its prompts and recommendations are reproducible, replicable, and consistent across diverse interactions and contexts. Achieving this ensures that healthcare providers can trust the model's outputs consistently, thereby streamlining their decision-making processes [50]. This enhanced reliability not only instills confidence in the LLM's capabilities, but also translates into more efficient and effective patient care, as clinicians can rely on the model to provide accurate and consistent guidance in various medical scenarios.

Although not the main focus of this study, it was evident that ChatGPT tended to generate lengthy and verbose responses, whereas Gemini provided more concise ones. In a clinical environment, where healthcare providers are frequently pressed for time and efficiency is paramount, the ability to deliver concise responses can be highly beneficial.

Given that every minute counts in such settings, the succinctness of Gemini's responses may offer a practical advantage, enabling healthcare professionals to quickly grasp essential information without unnecessary verbosity. In the high-pressure setting of the emergency department (ED), where healthcare providers are often inundated with urgent cases, the rapid delivery of concise, accurate information would be especially beneficial. Gemini's capability to provide succinct responses can support emergency providers in quickly assessing and prioritizing cases, thus enabling efficient resource allocation and expediting critical interventions.

ED applications of LLMs thus far have mostly focused on evaluating ChatGPT's ability to triage and diagnose. Berg et al. [48] examined ChatGPT's capacity to generate differential diagnoses, concluding that, while it can aid clinicians, its inconsistent responses limit its potential to replace clinical judgment. Meanwhile, Fraser et al. [51] compared ChatGPT-3.5 and ChatGPT-4 in triage and diagnosis, finding that ChatGPT-3.5 possessed high diagnostic accuracy but insufficient triage abilities. ChatGPT-4 showed improved triaging capability but lower diagnostic accuracy. The authors advised against unsupervised patient use and advocated for efforts to improve diagnostic and triage accuracy. Further studies highlighting ChatGPT's potential in emergency medicine include its role in suggesting diagnostic imaging [52] and providing diagnostic recommendations based on electrocardiography data [53]. Given that hand injuries are often first seen and evaluated in the ED [1,23,54], equipping emergency providers with resources like LLMs can help expedite triage and diagnostic workup.

The potential of LLMs to expedite diagnostic workup and management presents a promising solution for supporting emergency and primary care providers in managing hand injuries rapidly while waiting for a hand surgeon. This application could be especially advantageous in rural or underserved areas lacking on-site hand specialists. LLMs can empower these frontline providers to initiate diagnostic processes and treatment strategies, serving as a valuable resource until a hand specialist can evaluate the patient. This concept of using LLMs as a specialty consult has been previously discussed in the literature [50,51,55,56]. However, it is important to note that, while LLMs can support providers and bridge the gap between the initial presentation and hand surgery evaluation, they should not be used to replace an actual consultation with a hand specialist [1,11,13,24].

4.1. Ethical Considerations

As the integration of LLMs into medical practice becomes increasingly prevalent, it is imperative to address the ethical considerations and limitations associated with their use. While LLMs offer immense potential to enhance patient care and medical decision making in hand surgery, they also pose ethical challenges that necessitate careful attention. Upholding ethical principles, such as autonomy, beneficence, nonmaleficence, and justice, is paramount in the development and deployment of these models.

- **Autonomy:** Healthcare providers must ensure that LLMs respect a patient's autonomy by facilitating informed decision making and respecting their preferences and values, especially throughout the surgical process [57,58]. A patient's autonomy may be compromised if they are not adequately informed about the limitations, biases, and role of LLMs in their care [1,59].
- **Beneficence:** LLMs have the potential to significantly benefit patient care by providing timely and accurate information that can empower both healthcare professionals and patients to make more informed decisions. However, the implementation of LLMs must be guided by a commitment to maximizing these potential benefits while minimizing harm [48,59].
- **Nonmaleficence:** While LLMs can offer valuable assistance, they also carry inherent risks, including the potential for errors, biases, and misinformation [1,3,11,48]. Healthcare providers must critically evaluate and verify LLM-generated recommendations. Additionally, measures should be in place to mitigate the risk of LLMs propagating misinformation or perpetuating bias and healthcare disparities [57,59]. This entails the

ongoing monitoring and evaluation of LLM performance, as well as efforts to address any identified issues or limitations.

- Justice: The fair distribution of resources requires equitable access to this technology and its benefits [2]. Failing to address LLM bias and disparities in LLM utilization could exacerbate existing inequities in healthcare access and outcomes [1,2,52,57]. Therefore, it is imperative for healthcare systems to implement policies and initiatives aimed at promoting equitable access to LLM technology.

By paying careful attention to issues such as data privacy, transparency in decision-making processes, liability, and the mitigation of biases, healthcare can navigate the integration of LLMs in a manner that prioritizes ethical integrity. By proactively addressing these ethical considerations, healthcare can harness the full potential of LLMs while safeguarding patient autonomy, well-being, and justice in medical practice.

4.2. Limitations

We acknowledge multiple limitations to the study and the generalizability of its results. The clinical scenarios depicted in this study adhered closely to textbook examples. This prompts the question: how would these models fare when faced with less straightforward vignettes? Real-life patient encounters frequently involve atypical or complex presentations, which may diverge from the expected norms. Models trained solely on textbook-like cases may struggle to accurately interpret and respond to the complexities inherent in real-world medical practice. Evaluating these models' adaptability to a wide range of clinical scenarios is essential for their effectiveness. Failing to do so risks undermining their reliability and applicability in real-world healthcare settings, potentially compromising patient care outcomes. Therefore, future research endeavors should prioritize testing these models against a wider range of clinical vignettes to comprehensively assess their real-world utility and identify areas for improvement.

Although this study included 68 unique patient vignettes covering 12 classification systems, it is by no means comprehensive. While these vignettes offer valuable insights into the classification abilities of the models under examination, they may not fully encapsulate the diverse spectrum of hand injuries encountered in clinical practice. As such, the study's findings should be interpreted within the context of its inherent constraints. Despite these challenges, our study serves as a starting point for future investigations to delve into these nuances and to advance our understanding of LLM performance in real-world clinical practice.

Furthermore, the success of using LLMs in medicine depends entirely on their ability to provide accurate and reliable information. The speed at which an LLM can respond to queries becomes irrelevant if it offers incorrect and potentially harmful recommendations. Misinformation can lead to adverse patient outcomes and the erosion of trust in technology and healthcare providers [9,57,60]. As previously mentioned, we acknowledge that neither ChatGPT nor Gemini was designed specifically with medical applications in mind. The datasets on which these models were trained likely lack significant medical information and data. This inherent limitation underscores the importance of further refining LLMs specifically for healthcare applications in order to mitigate such shortcomings.

4.3. Future Research and Next Steps

Moving forward, it is imperative to address the limitations highlighted in this study to maximize the potential of LLMs in healthcare. While ChatGPT and Gemini have exhibited promising capabilities in classifying hand injuries and offering management recommendations, there remains a need for further research to refine and enhance their performance for real-world clinical applications. One crucial area of focus for future investigation involves expanding the training datasets of these LLMs with more comprehensive medical information and clinical data. This would allow these models to better navigate the intricacies of medical decision making. Additionally, future studies should aim to evaluate the performance of ChatGPT, Gemini, and other LLMs in classifying atypical or complex hand

injuries, as well as extending their assessment to encompass a broader spectrum of injuries and medical conditions. Most current studies focus exclusively on ChatGPT, but with Gemini's superior performance in hand injury classification, further studies examining this model's ability are warranted. By systematically examining LLM performance in diverse clinical scenarios, researchers can identify areas for improvement and tailor these models to address the specific challenges encountered in medical practice.

Furthermore, the advancement of LLMs in medicine necessitates ongoing research and development. By continually refining and validating AI tools like ChatGPT and Gemini, healthcare professionals can harness their full potential as invaluable resources in clinical practice. Further investigation into the use and performance of additional AI tools like RAG, especially when used in conjunction with LLMs, is indicated. Through collaborative interdisciplinary efforts between developers, healthcare institutions, and medical professionals, the development of robust and reliable medical-focused LLMs can pave the way for a new era of personalized and efficient healthcare delivery.

5. Conclusions

This study evaluates the performance of ChatGPT and Gemini in classifying hand injuries and suggesting management. While both models show potential, Gemini generally performs better than ChatGPT in classification, although not currently at a suitable level for current use. For treatment recommendations, ChatGPT leans towards recommending surgical intervention more readily, albeit with lower specificity than Gemini. These findings stress the need to carefully weigh the strengths and limitations of different LLMs when incorporating them into clinical practice. Both ChatGPT and Gemini hold promise as valuable resources for hand surgeons. This potential is expected to translate into enhanced diagnostic accuracy and treatment decisions, ultimately improving patient outcomes. However, further development and research are necessary to ensure the reliability of these models.

Author Contributions: Conceptualization, S.M.P. and A.J.F.; methodology, S.M.P., S.B., C.A.G.-C. and S.A.H.; software, S.M.P., S.B. and A.J.F.; validation, S.M.P. and S.B.; formal analysis, S.M.P.; investigation, S.M.P.; resources, NA; data curation, S.M.P., S.B., C.A.G.-C. and S.A.H.; writing—original draft preparation, S.M.P., S.B., C.A.G.-C. and S.A.H.; writing—review and editing, S.M.P., S.B., C.A.G.-C., S.A.H. and A.J.F.; visualization, S.M.P. and A.J.F.; supervision, A.J.F.; project administration, A.J.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval was not applicable for this study since no patient information was used. Patient scenarios were fictionalized.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors acknowledge the use of ChatGPT in text editing and the creation of icons for Figure 4. The authors assume full responsibility for the content of this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Miller, R.; Farnebo, S.; Horwitz, M.D. Insights and trends review: Artificial intelligence in hand surgery. *J. Hand Surg. Eur. Vol.* **2023**, *48*, 396–403. [[CrossRef](#)] [[PubMed](#)]
2. Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. [[CrossRef](#)] [[PubMed](#)]
3. Dave, T.; Athaluri, S.A.; Singh, S. ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front. Artif. Intell.* **2023**, *6*, 1169595. [[CrossRef](#)] [[PubMed](#)]
4. Ulusoy, I.; Yilmaz, M.; Kivrak, A. How Efficient Is ChatGPT in Accessing Accurate and Quality Health-Related Information? *Cureus* **2023**, *15*, e46662. [[CrossRef](#)] [[PubMed](#)]

5. Mikolov, T.; Karafiát, M.; Burget, L.; Cernocký, J.; Khudanpur, S. Recurrent neural network based language model. In *Interspeech*; ISCA: Chiba, Japan, 2010; pp. 1045–1048.
6. Jin, Z. Analysis of the Technical Principles of ChatGPT and Prospects for Pre-trained Large Models. In Proceedings of the 2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), Chongqing, China, 26–28 May 2023; pp. 1755–1758.
7. Google. Gemini. Available online: <https://gemini.google.com/app> (accessed on 10 March 2024).
8. OpenAI. ChatGPT. Available online: <https://chat.openai.com/chat> (accessed on 10 March 2024).
9. Abi-Rafeh, J.; Xu, H.H.; Kazan, R.; Tevlin, R.; Furnas, H. Large Language Models and Artificial Intelligence: A Primer for Plastic Surgeons on the Demonstrated & Potential Applications, Promises, and Limitations of ChatGPT. *Aesthet. Surg. J.* **2023**, *44*, 329–343. [[CrossRef](#)] [[PubMed](#)]
10. 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 Digit. Health* **2023**, *2*, e0000198. [[CrossRef](#)] [[PubMed](#)]
11. Ghanem, D.; Nassar, J.; El Bachour, J.; Hanna, T. ChatGPT Earns American Board Certification in Hand Surgery. *Hand Surg. Rehabil.* **2024**, 101688. [[CrossRef](#)] [[PubMed](#)]
12. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11*, 887. [[CrossRef](#)] [[PubMed](#)]
13. Keller, M.; Gubeli, A.; Thieringer, F.; Honigmann, P. Artificial intelligence in patient-specific hand surgery: A scoping review of literature. *Int. J. Comput. Assist. Radiol. Surg.* **2023**, *18*, 1393–1403. [[CrossRef](#)]
14. Gummesson, C.; Ward, M.M.; Atroshi, I. The shortened disabilities of the arm, shoulder and hand questionnaire (Quick DASH): Validity and reliability based on responses within the full-length DASH. *BMC Musculoskelet. Disord.* **2006**, *7*, 1–7. [[CrossRef](#)]
15. Poerbodipoero, S.; Steultjens, M.; Van der Beek, A.; Dekker, J. Pain, disability in daily activities and work participation in patients with traumatic hand injury. *Br. J. Hand Ther.* **2007**, *12*, 40–47. [[CrossRef](#)]
16. Schier, J.S.; Chan, J. Changes in life roles after hand injury. *J. Hand Ther.* **2007**, *20*, 57–69. [[CrossRef](#)] [[PubMed](#)]
17. Smith, M.E.; Auchincloss, J.M.; Ali, M.S. Causes and consequences of hand injury. *J. Hand Surg. Br.* **1985**, *10*, 288–292. [[CrossRef](#)] [[PubMed](#)]
18. Angly, B.; Constantinescu, M.A.; Kreutziger, J.; Juon, B.H.; Vögelin, E. Early versus delayed surgical treatment in open hand injuries: A paradigm revisited. *World J. Surg.* **2012**, *36*, 826–829. [[CrossRef](#)]
19. del Pinal, F. Severe mutilating injuries to the hand: Guidelines for organizing the chaos. *J. Plast. Reconstr. Aesthet. Surg.* **2007**, *60*, 816–827. [[CrossRef](#)] [[PubMed](#)]
20. Gustilo, R.B.; Anderson, J.T. Prevention of infection in the treatment of one thousand and twenty-five open fractures of long bones: Retrospective and prospective analyses. *J. Bone Joint Surg. Am.* **1976**, *58*, 453–458. [[CrossRef](#)] [[PubMed](#)]
21. Salazar Botero, S.; Hidalgo Diaz, J.J.; Benaïda, A.; Collon, S.; Facca, S.; Liverneaux, P.A. Review of Acute Traumatic Closed Mallet Finger Injuries in Adults. *Arch. Plast. Surg.* **2016**, *43*, 134–144. [[CrossRef](#)]
22. Wong, K.; von Schroeder, H.P. Delays and Poor Management of Scaphoid Fractures: Factors Contributing to Nonunion. *J. Hand Surg.* **2011**, *36*, 1471–1474. [[CrossRef](#)]
23. Yoong, P.; Johnson, C.A.; Yoong, E.; Chojnowski, A. Four hand injuries not to miss: Avoiding pitfalls in the emergency department. *Eur. J. Emerg. Med.* **2011**, *18*, 186–191. [[CrossRef](#)] [[PubMed](#)]
24. Leybold, T.; Schäfer, B.; Boos, A.; Beier, J.P. Can AI Think Like a Plastic Surgeon? Evaluating GPT-4's Clinical Judgment in Reconstructive Procedures of the Upper Extremity. *Plast. Reconstr. Surg. Glob. Open* **2023**, *11*, e5471. [[CrossRef](#)]
25. Crook, B.S.; Park, C.N.; Hurley, E.T.; Richard, M.J.; Pidgeon, T.S. Evaluation of Online Artificial Intelligence-Generated Information on Common Hand Procedures. *J. Hand Surg. Am.* **2023**, *48*, 1122–1127. [[CrossRef](#)] [[PubMed](#)]
26. Seth, I.; Xie, Y.; Rodwell, A.; Gracias, D.; Bulloch, G.; Hunter-Smith, D.J.; Rozen, W.M. Exploring the Role of a Large Language Model on Carpal Tunnel Syndrome Management: An Observation Study of ChatGPT. *J. Hand Surg. Am.* **2023**, *48*, 1025–1033. [[CrossRef](#)] [[PubMed](#)]
27. Al Rawi, Z.M.; Kirby, B.J.; Albrecht, P.A.; Nuelle, J.A.V.; London, D.A. Experimenting With the New Frontier: Artificial Intelligence-Powered Chat Bots in Hand Surgery. *Hand* **2024**, 15589447241238372. [[CrossRef](#)] [[PubMed](#)]
28. Cooney, W.P., 3rd. Scaphoid fractures: Current treatments and techniques. *Instr. Course Lect.* **2003**, *52*, 197–208. [[PubMed](#)]
29. Cooney, W.P.; Dobyns, J.H.; Linscheid, R.L. Fractures of the scaphoid: A rational approach to management. *Clin. Orthop. Relat. Res.* **1980**, *149*, 90–97. [[CrossRef](#)]
30. Eaton, R.G.; Malerich, M.M. Volar plate arthroplasty of the proximal interphalangeal joint: A review of ten years' experience. *J. Hand Surg. Am.* **1980**, *5*, 260–268. [[CrossRef](#)] [[PubMed](#)]
31. Geissler, W.B. Arthroscopic management of scapholunate instability. *J. Wrist Surg.* **2013**, *2*, 129–135. [[CrossRef](#)] [[PubMed](#)]
32. Green, D.P.; O'Brien, E.T. Fractures of the thumb metacarpal. *South. Med. J.* **1972**, *65*, 807–814. [[CrossRef](#)] [[PubMed](#)]
33. Gustilo, R.B.; Mendoza, R.M.; Williams, D.N. Problems in the management of type III (severe) open fractures: A new classification of type III open fractures. *J. Trauma.* **1984**, *24*, 742–746. [[CrossRef](#)] [[PubMed](#)]
34. Herbert, T.J.; Fisher, W.E. Management of the fractured scaphoid using a new bone screw. *J. Bone Joint Surg. Br.* **1984**, *66*, 114–123. [[CrossRef](#)]

35. Hintermann, B.; Holzach, P.J.; Schütz, M.; Matter, P. Skier's thumb--the significance of bony injuries. *Am. J. Sports Med.* **1993**, *21*, 800–804. [[CrossRef](#)] [[PubMed](#)]
36. Kleinert, H.E.; Verdan, C. Report of the Committee on Tendon Injuries. *J. Hand Surg.* **1983**, *8*, 794–798. [[CrossRef](#)] [[PubMed](#)]
37. Leddy, J.P.; Packer, J.W. Avulsion of the profundus tendon insertion in athletes. *J. Hand Surg. Am.* **1977**, *2*, 66–69. [[CrossRef](#)]
38. Lichtman, D.M.; Pientka, W.F., 2nd; Bain, G.I. Kienböck Disease: A New Algorithm for the 21st Century. *J. Wrist Surg.* **2017**, *6*, 2–10. [[CrossRef](#)] [[PubMed](#)]
39. Mayfield, J.K.; Johnson, R.P.; Kilcoyne, R.K. Carpal dislocations: Pathomechanics and progressive perilunar instability. *J. Hand Surg. Am.* **1980**, *5*, 226–241. [[CrossRef](#)]
40. Carlà, M.M.; Gambini, G.; Baldascino, A.; Boselli, F.; Giannuzzi, F.; Margollicci, F.; Rizzo, S. Large language models as assistance for glaucoma surgical cases: A ChatGPT vs. Google Gemini comparison. *Graefes Arch. Clin. Exp. Ophthalmol.* **2024**. [[CrossRef](#)] [[PubMed](#)]
41. Carlà, M.M.; Gambini, G.; Baldascino, A.; Giannuzzi, F.; Boselli, F.; Crincoli, E.; D'Onofrio, N.C.; Rizzo, S. Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. *Br. J. Ophthalmol.* **2024**. [[CrossRef](#)]
42. Koga, S.; Martin, N.B.; Dickson, D.W. Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. *Brain Pathol.* **2023**, *34*, e13207. [[CrossRef](#)] [[PubMed](#)]
43. Kumari, A.; Kumari, A.; Singh, A.; Singh, S.K.; Juhi, A.; Dhanvijay, A.K.D.; Pinjar, M.J.; Mondal, H. Large Language Models in Hematology Case Solving: A Comparative Study of ChatGPT-3.5, Google Bard, and Microsoft Bing. *Cureus* **2023**, *15*, e43861. [[CrossRef](#)]
44. Lim, Z.W.; Pushpanathan, K.; Yew, S.M.E.; Lai, Y.; Sun, C.H.; Lam, J.S.H.; Chen, D.Z.; Goh, J.H.L.; Tan, M.C.J.; Sheng, B.; et al. Benchmarking large language models' performances for myopia care: A comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine* **2023**, *95*, 104770. [[CrossRef](#)]
45. Rahsepar, A.A.; Tavakoli, N.; Kim, G.H.J.; Hassani, C.; Abtin, F.; Bedayat, A. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. *Radiology* **2023**, *307*, e230922. [[CrossRef](#)] [[PubMed](#)]
46. Gan, R.K.; Ogbodo, J.C.; Wee, Y.Z.; Gan, A.Z.; González, P.A. Performance of Google bard and ChatGPT in mass casualty incidents triage. *Am. J. Emerg. Med.* **2024**, *75*, 72–78. [[CrossRef](#)] [[PubMed](#)]
47. Zúñiga Salazar, G.; Zúñiga, D.; Vindel, C.L.; Yoong, A.M.; Hincapie, S.; Zúñiga, A.B.; Zúñiga, P.; Salazar, E.; Zúñiga, B. Efficacy of AI Chats to Determine an Emergency: A Comparison Between OpenAI's ChatGPT, Google Bard, and Microsoft Bing AI Chat. *Cureus* **2023**, *15*, e45473. [[CrossRef](#)]
48. Berg, H.T.; van Bakel, B.; van de Wouw, L.; Jie, K.E.; Schipper, A.; Jansen, H.; O'Connor, R.D.; van Ginneken, B.; Kurstjens, S. ChatGPT and Generating a Differential Diagnosis Early in an Emergency Department Presentation. *Ann. Emerg. Med.* **2024**, *83*, 83–86. [[CrossRef](#)] [[PubMed](#)]
49. Franc, J.M.; Cheng, L.; Hart, A.; Hata, R.; Hertelendy, A. Repeatability, reproducibility, and diagnostic accuracy of a commercial large language model (ChatGPT) to perform emergency department triage using the Canadian triage and acuity scale. *Cjem* **2024**, *26*, 40–46. [[CrossRef](#)] [[PubMed](#)]
50. Funk, P.F.; Hoch, C.C.; Knoedler, S.; Knoedler, L.; Cotofana, S.; Sofu, G.; Bashiri Dezfouli, A.; Wollenberg, B.; Guntinas-Lichius, O.; Alfertshofer, M. ChatGPT's Response Consistency: A Study on Repeated Queries of Medical Examination Questions. *Eur. J. Investig. Health Psychol. Educ.* **2024**, *14*, 657–668. [[CrossRef](#)]
51. Fraser, H.; Crossland, D.; Bacher, I.; Ranney, M.; Madsen, T.; Hilliard, R. Comparison of Diagnostic and Triage Accuracy of Ada Health and WebMD Symptom Checkers, ChatGPT, and Physicians for Patients in an Emergency Department: Clinical Data Analysis Study. *JMIR Mhealth Uhealth* **2023**, *11*, e49995. [[CrossRef](#)] [[PubMed](#)]
52. Barash, Y.; Klang, E.; Konen, E.; Sorin, V. ChatGPT-4 Assistance in Optimizing Emergency Department Radiology Referrals and Imaging Selection. *J. Am. Coll. Radiol.* **2023**, *20*, 998–1003. [[CrossRef](#)] [[PubMed](#)]
53. Günay, S.; Öztürk, A.; Özerol, H.; Yiğit, Y.; Erenler, A.K. Comparison of emergency medicine specialist, cardiologist, and chat-GPT in electrocardiography assessment. *Am. J. Emerg. Med.* **2024**, *80*, 51–60. [[CrossRef](#)]
54. van Leerdam, R.H.; Krijnen, P.; Panneman, M.J.; Schipper, I.B. Incidence and treatment of hand and wrist injuries in Dutch emergency departments. *Eur. J. Trauma. Emerg. Surg.* **2022**, *48*, 4327–4332. [[CrossRef](#)]
55. Rizwan, A.; Sadiq, T. The Use of AI in Diagnosing Diseases and Providing Management Plans: A Consultation on Cardiovascular Disorders With ChatGPT. *Cureus* **2023**, *15*, e43106. [[CrossRef](#)] [[PubMed](#)]
56. Sun, Y.X.; Li, Z.M.; Huang, J.Z.; Yu, N.Z.; Long, X. GPT-4: The Future of Cosmetic Procedure Consultation? *Aesthet. Surg. J.* **2023**, *43*, NP670–NP672. [[CrossRef](#)] [[PubMed](#)]
57. Oleck, N.C.; Naga, H.I.; Nichols, D.S.; Morris, M.X.; Dhingra, B.; Patel, A. Navigating the Ethical Landmines of ChatGPT: Implications of Intelligent Chatbots in Plastic Surgery Clinical Practice. *Plast. Reconstr. Surg. Glob. Open* **2023**, *11*, e5290. [[CrossRef](#)] [[PubMed](#)]
58. Pressman, S.M.; Borna, S.; Gomez-Cabello, C.A.; Haider, S.A.; Haider, C.; Forte, A.J. AI and Ethics: A Systematic Review of the Ethical Considerations of Large Language Model Use in Surgery Research. *Healthcare* **2024**, *12*, 825. [[CrossRef](#)] [[PubMed](#)]

-
59. Keskinbora, K.H. Medical ethics considerations on artificial intelligence. *J. Clin. Neurosci.* **2019**, *64*, 277–282. [[CrossRef](#)]
 60. Li, W.; Zhang, Y.; Chen, F. ChatGPT in Colorectal Surgery: A Promising Tool or a Passing Fad? *Ann. Biomed. Eng.* **2023**, *51*, 1892–1897. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.