Large language models in upper limb surgery: A narrative review

Catherine Simister, Kyle Lam, Andrew Yiu and James M Kinross

Catherine Simister is a final year medical student at Imperial College London with an interest in surgery.

Kyle Lam is a NIHR Academic Clinical Fellow and Registrar in General Surgery at Imperial College London. He completed a PhD in AI for surgical performance assessment. He has wide ranging research interests within surgical AI including computer vision, large language models, and challenges in clinical implementation.

Adoption of large language models (LLMs) has grown exponentially throughout society, and the most well-known of these is ChatGPT (OpenAI, San Francisco, CA, USA). Since in its launch in 2022, it now averages 100 million weekly users[1,2]. LLMs represent a form of artificial intelligence (AI) that are capable of understanding and generating human-like text[3,4].

While the workings of LLMs are more complex than this description, in its simplest form, these models work by predicting the next word in a phrase. If a human is asked to complete the sentence, 'Humpty Dumpty sat on a', they would most likely say 'wall' and LLMs perform a similar task in reading the input text and predicting what word is most likely to follow it. The input or 'prompt' (Table 1) to a LLM can similarly turn this next-word prediction mechanism into a question answering machine, for example, 'The nerve innervating the biceps brachii muscle is…'. When confronted by this, the LLM must predict what word would come next through the generation of probabilities for possible next words based on patterns from the data it was trained on and the selection of the highest probability word. Here, the model might output a probability of 95% for musculocutaneous, 2% for radial, and 1% for median and therefore select musculocutaneous.

The power of these models lies in the fact that LLMs learn this next word prediction through the recognition of patterns in enormous quantities of textual data. Commonly available LLMs on the market (Table 2) are likely to have been trained across the majority of textual data available on the internet including internet forum discussions and Wikipedia via a process called self-supervised learning. Differing from supervised learning, where the model would learn that a picture was that of a cat because it had been annotated by a human as a cat, in self-supervised learning

Term	Definition
Artificial Intelligence	The science and engineering of making intelligent machines.
Machine Learning	The field of study that gives computers the ability to learn without explicitly being programmed.
Large Language Model (LLM)	AI systems capable of understanding and generating human language by processing vast amounts of text data.
Generative AI	A type of artificial intelligence that can create new content ranging from text, images, or video based on data it has learnt from.
Parameter	A variable learnt from the data during the training process of an AI model which is subsequently used to make predictions on new data.
Supervised learning	A model that is trained on labelled inputs to classify data or predict outcomes.
Unsupervised learning	A model that learns patterns exclusively from unlabelled data.
Semi-supervised learning	A model that learns from a small portion of labelled data and lots of unlabelled data to train a predictive model.
Fine-tuning	The process of taking an AI model and further training it on a smaller targeted dataset. The aims are for the model to keep the original capabilities of the initial model while adapting it to more specialised use cases.
Prompt	In the context of a LLM, this is the input or set of instructions given to the LLM.

Table 1: Definitions of common terms.

Andrew Yiu is a General Surgery Registrar in Surrey and Honorary Research Fellow at Imperial College London. He has an interest in digital surgery with a focus on improving patient outcomes using operative data, particularly surgical video.

James Kinross is a Reader in Colorectal Surgery and a Consultant Surgeon at Imperial College London. His clinical interest is in robotic surgery and minimally invasive surgery for colorectal cancer. He performs translational research in the fields of early colorectal cancer detection and prevention and in surgical technology transfer.

the model would guess 'Dumpty' if 'Humpty' was inputted and then compare this prediction with the actual word contained within the training data. If correct, the model then updates the variables within the model or 'parameters' so that it is more likely to predict this the next time. The model subsequently looks at 'Humpty Dumpty' and tries to predict the third word in the sentence by repeating the same process.

The speed of societal AI adoption has been driven in part by the increase in computing power for the training of generative AI based models, which is climbing fourfold each year, and by novel developments in AI architectures. AI capable of creating original content, such as these models, are termed 'generative AI' and can output text, images, video, and audio, from minimal input data[5,6]. Generative AI has seen rapid uptake across multiple industries including the financial sector for report writing and within retail for online chatbot services[7] and this has been reflected within its growth within the market having been estimated to be worth over $1 trillion by 2032. It is therefore little surprise that healthcare has also been a proposed target for LLMs[8,9]. This review sets out prospective uses and benefits of LLMs within healthcare, with a focus on upper limb surgery, describes envisaged challenges to clinical translation, and future targets for research.

LLMs encode clinical knowledge

LLMs have been found to encode significant clinical knowledge, with ChatGPT 3.5 either meeting or achieving close to the pass mark in all three steps of the US Medical Licensing Exam (USMLE)[10-12]. Within postgraduate examinations, LLMs have also passed the Member of the Royal College of Surgeons (MRCS) exam, scoring more than 85%[12]. LLMs, such as MedPaLM and PubMedGPT, which have been specifically fine-tuned on medical datasets may enhance the trustworthiness of 'out-of-the-box' models and may demonstrate further promise for the use of LLMs within clinical practice[13,14].

One proposed clinical application for LLMs within the literature are for LLMs to act as co-pilots for diagnosis and management. Within upper limb surgery, Google's Gemini correctly classified 70% of hand injuries from a series of textbook vignettes[15] and Daher *et al.* found ChatGPT 3.5 correctly diagnosed 93% of shoulder and elbow complaints, but relied heavily on MRI reports[16]. ChatGPT's diagnostic success was, however, not replicated when presented with more complex cases, such as patients with comorbidities or uncommon injuries[16,17]. For example, ChatGPT 3.5 was unable to diagnose an atypical radial nerve injury where the sensory impairment differed from the conventional pattern[17]. Whilst models themselves are limited by the data they are trained upon, they are also limited by the input or 'prompt' they are given. Limitations of these models will likely be improved upon with future technological iterations which will likely see an exponential growth in model size.

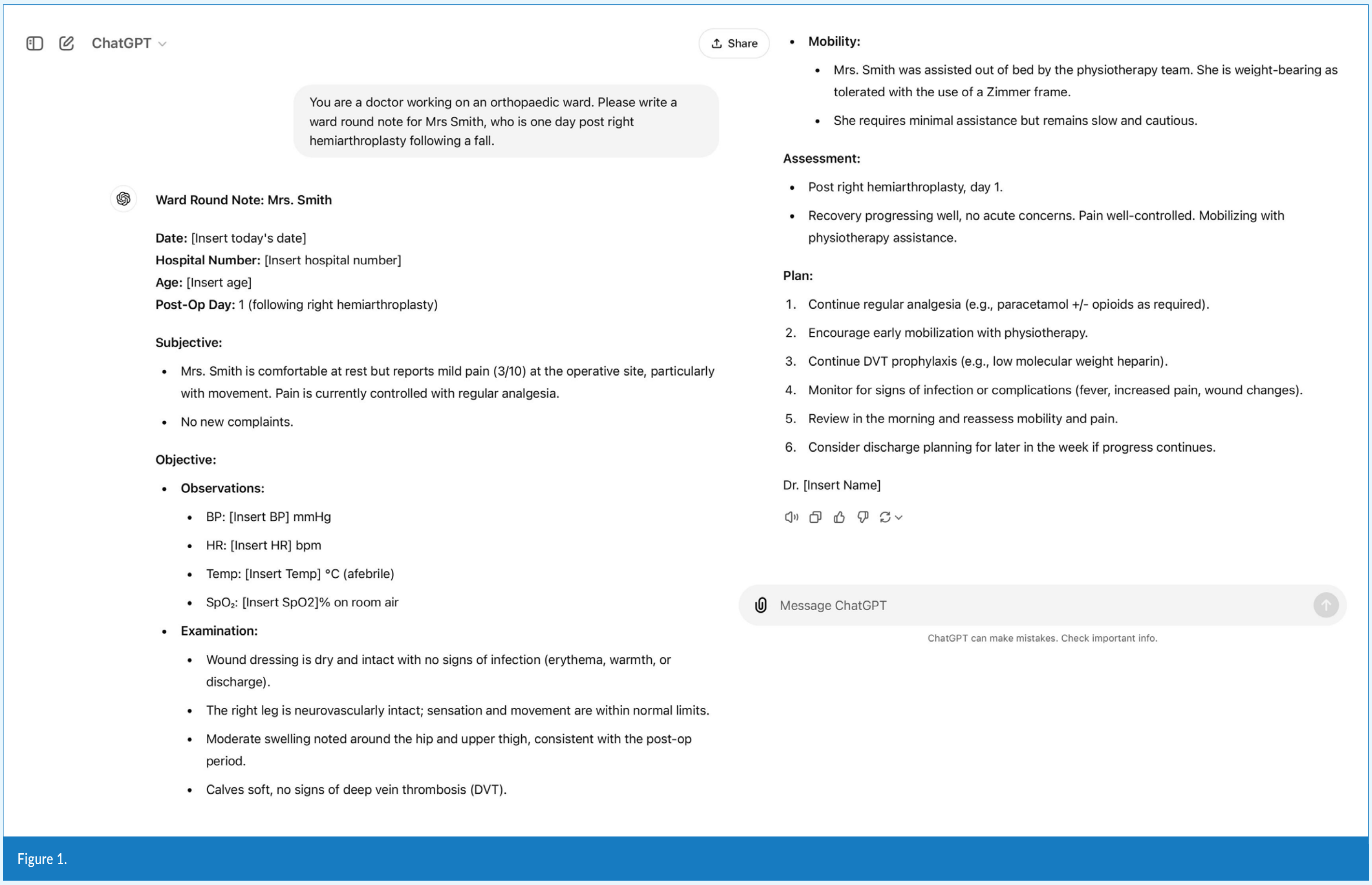
LLMs for improved clinical efficiency

LLMs have also been proposed to improve clinical efficiency allowing time to be more effectively dedicated to direct patient care[18]. One potential application is within triage, where LLMs could facilitate clinical decision-making and automatically streamline patients to the correct services at presentation[19-20]. Newer iterations of LLMs, which allow processing of multimodal input, such as radiographic images, open significant avenues for their wider use, especially within orthopaedic surgery[21,22].

Beyond direct clinical use, the capability of LLMs to automatically generate text often indistinguishable from that authored by humans could see them alleviating the high administrative burden of clerical work for junior doctors, who are estimated to spend approximately 50% of their time completing these tasks[23], (Figure 1). >>

Large Language Model	Creator	Training Data	Parameters	Features	Access
ChatGPT	OpenAI	Not disclosed	Estimated 1.76 trillion (GPT4)	Multimodal LLM with improved reasoning compared to ChatGPT3.5	Free to access basic model, monthly subscription for ChatGPT plus
Gemini	Google	Not disclosed	Ultra (1.5T), Pro (350B), Nano (100B)	Multimodal LLM with different size models for various use cases	Free tier, paid tiers for advanced features
Claude	Anthropic	Not disclosed	175 billion (Claude 2)	LLM with strong capabilities in analysis, coding, and task completion	Free tier with limitations, paid plans for full access
LLaMA	Meta	1.4 trillion tokens	7B, 13B, 33B, 65B versions	Open-source LLM, adaptable for various tasks	Free, open-source (with license agreement)
MedPaLM	Google	Medical datasets and general data	Not disclosed	Specialised LLM for medical knowledge and tasks	Research access only, not publicly available
MedAlpaca	Stanford	LLaMA base fine tuned on medical datasets	7B, 13B versions	Open-source medical LLM based on LLaMA	Free, open-source

Table 2: Examples of currently available large language models.



You are a doctor working on an orthopaedic ward. Please write a ward round note for Mrs Smith, who is one day post right hemiarthroplasty following a fall.

Ward Round Note: Mrs. Smith

Date: [Insert today's date]
Hospital Number: [Insert hospital number]
Age: [Insert age]
Post-Op Day: 1 (following right hemiarthroplasty)

Subjective:

- Mrs. Smith is comfortable at rest but reports mild pain (3/10) at the operative site, particularly with movement. Pain is currently controlled with regular analgesia.
- No new complaints.

Objective:

- **Observations:**
  - BP: [Insert BP] mmHg
  - HR: [Insert HR] bpm
  - Temp: [Insert Temp] °C (afebrile)
  - SpO$_2$: [Insert SpO2]% on room air
- **Examination:**
  - Wound dressing is dry and intact with no signs of infection (erythema, warmth, or discharge).
  - The right leg is neurovascularly intact; sensation and movement are within normal limits.
  - Moderate swelling noted around the hip and upper thigh, consistent with the post-op period.
  - Calves soft, no signs of deep vein thrombosis (DVT).
- **Mobility:**
  - Mrs. Smith was assisted out of bed by the physiotherapy team. She is weight-bearing as tolerated with the use of a Zimmer frame.
  - She requires minimal assistance but remains slow and cautious.

Assessment:

- Post right hemiarthroplasty, day 1.
- Recovery progressing well, no acute concerns. Pain well-controlled. Mobilizing with physiotherapy assistance.

Plan:

1. Continue regular analgesia (e.g., paracetamol +/- opioids as required).
2. Encourage early mobilization with physiotherapy.
3. Continue DVT prophylaxis (e.g., low molecular weight heparin).
4. Monitor for signs of infection or complications (fever, increased pain, wound changes).
5. Review in the morning and reassess mobility and pain.
6. Consider discharge planning for later in the week if progress continues.

Dr. [Insert Name]

Figure 1.

Patel and Lam found that ChatGPT was able to produce a discharge summary when provided with a short clinical vignette, suggesting LLMs may be useful in reducing the time taken to complete these tasks[24]. However, such applications must be balanced with the need to maintain patient confidentiality, which can limit the capability of LLMs both within the training and inference process[25].

LLMs for education

There are significant challenges ahead of translation of LLMs for direct clinical applications, including quality assurance, trustworthiness, and security. Education, however, is a potentially low-risk, high-reward setting for the use of LLMs, which could breach the gap prior to their wider clinical use. For example, LLMs could assist with simulating patient-doctor scenarios, creating study plans, exploring concepts, and providing multiple choice question response explanations, (Figure 2), with few consequences in the case of algorithmic failure[26]. As previously discussed, LLMs can encode clinical knowledge, but within orthopaedic education, performance of LLMs has been variable. Lum *et al.*, concluded that ChatGPT could perform at a second-year resident's level in the Orthopaedic In-Training Examination (OITE), although another study found that ChatGPT was able to achieve a score above that of residents at any level[27.28]. When tested on exclusively upper limb questions, however, ChatGPT only achieved 45%, well below the expected standard[29]. This study, however, demonstrated the need for effective prompt engineering as the model was likely limited by the use of manual descriptions of images within the prompt rather than the input of the raw images themselves[29].

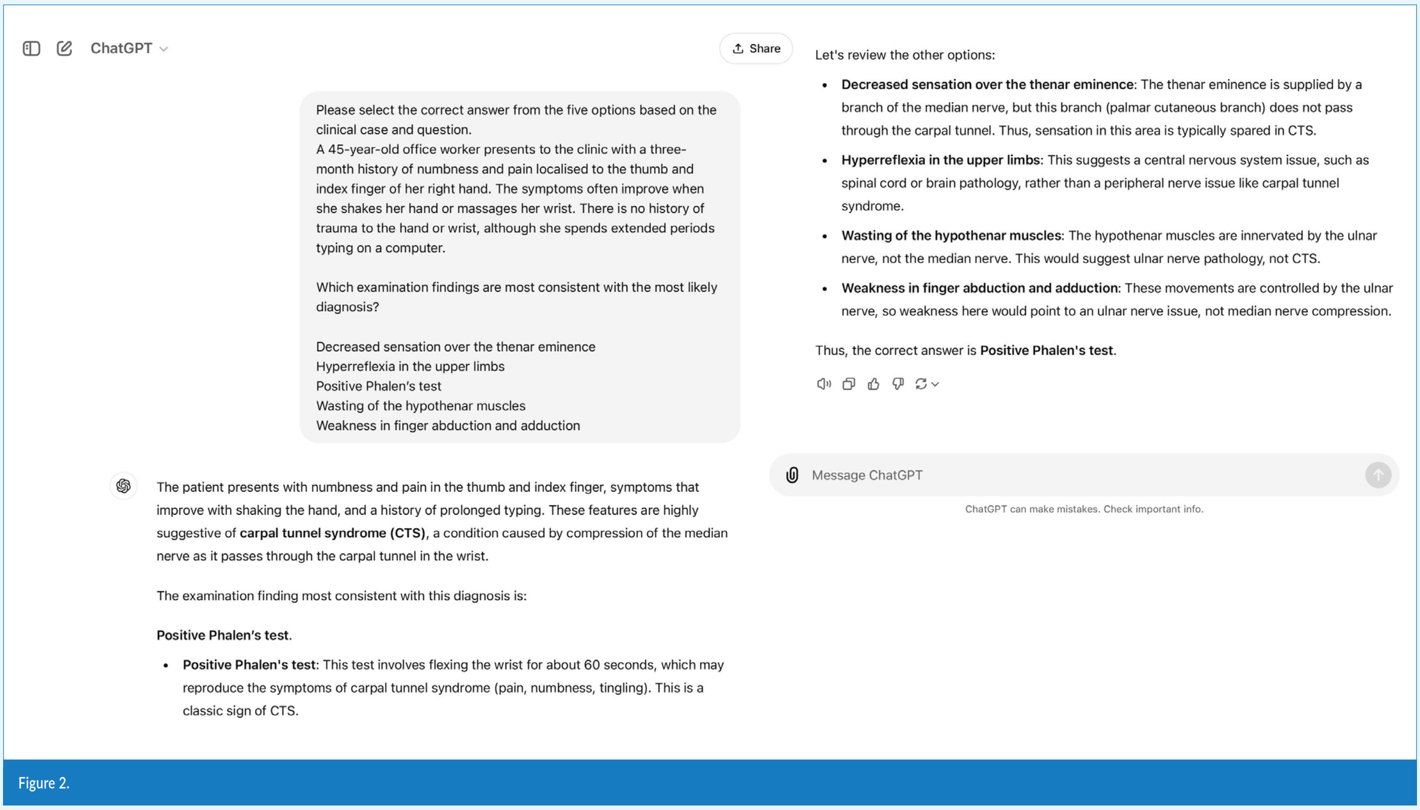
LLMs for academia

Academia also stands to benefit from the skills of LLMs, which could assist in multiple stages of research[30]. LLMs may assist in the brainstorming phase, as well as providing elaboration on existing ideas[31]. LLMs could shorten the notoriously time-consuming tasks of grant and manuscript writing, plus offer grammar checking for final drafts. Their ability to summarise text could be useful in writing abstracts, and the use of LLMs for this purpose has already been demonstrated, with Bisi *et al.*, finding the highest rates of AI-generated text within abstracts[30-32]. Yet the use of LLMs to generate academic text is controversial, raising questions over academic integrity[33,34]. Whilst guidelines are not standardised, most leading medical publishers do not accept LLMs as an author on scientific papers and require disclosure if LLMs have been used beyond grammar and spell-checking purposes. However, the challenge of identifying AI generated text remains[35]. The need to create methods to detect AI generated text and standardised guidelines between journals is therefore urgent.

LLMs for patients

Patients are also increasingly likely to utilise LLMs and there are several possible ways in which LLMs could alter the way patients interact with healthcare providers. The Topol Review, which explored future roles of digital technology in healthcare, cited telemedicine as a major advancement[36]. LLMs are likely to play a significant role in telemedicine. The now-extinct Babylon Health was used to assist in the initial triage, diagnosis and management of patients, but failed due to cost inefficiency and limited usefulness of its chatbot service[37]. There is significant potential for LLMs or 'chatbots' to assume a triage role, facilitating direction towards the most appropriate service or management.

It is not uncommon for patients to consult 'Dr Google' and with the ease of access, many may look to LLMs to answer questions about their health. ChatGPT's ability to answer common questions relating to



Please select the correct answer from the five options based on the clinical case and question.
A 45-year-old office worker presents to the clinic with a three-month history of numbness and pain localised to the thumb and index finger of her right hand. The symptoms often improve when she shakes her hand or massages her wrist. There is no history of trauma to the hand or wrist, although she spends extended periods typing on a computer.

Which examination findings are most consistent with the most likely diagnosis?

Decreased sensation over the thenar eminence
Hyperreflexia in the upper limbs
Positive Phalen's test
Wasting of the hypothenar muscles
Weakness in finger abduction and adduction

The patient presents with numbness and pain in the thumb and index finger, symptoms that improve with shaking the hand, and a history of prolonged typing. These features are highly suggestive of **carpal tunnel syndrome (CTS)**, a condition caused by compression of the median nerve as it passes through the carpal tunnel in the wrist.

The examination finding most consistent with this diagnosis is:

Positive Phalen's test.

- **Positive Phalen's test**: This test involves flexing the wrist for about 60 seconds, which may reproduce the symptoms of carpal tunnel syndrome (pain, numbness, tingling). This is a classic sign of CTS.

Let's review the other options:

- **Decreased sensation over the thenar eminence**: The thenar eminence is supplied by a branch of the median nerve, but this branch (palmar cutaneous branch) does not pass through the carpal tunnel. Thus, sensation in this area is typically spared in CTS.
- **Hyperreflexia in the upper limbs**: This suggests a central nervous system issue, such as spinal cord or brain pathology, rather than a peripheral nerve issue like carpal tunnel syndrome.
- **Wasting of the hypothenar muscles**: The hypothenar muscles are innervated by the ulnar nerve, not the median nerve. This would suggest ulnar nerve pathology, not CTS.
- **Weakness in finger abduction and adduction**: These movements are controlled by the ulnar nerve, so weakness here would point to an ulnar nerve issue, not median nerve compression.

Thus, the correct answer is **Positive Phalen's test**.

ChatGPT can make mistakes. Check important info.

Figure 2.

upper limb conditions has been explored extensively[38-40]. One study looking at questions relating to rotator cuff injuries, found that whilst answers were generally correct and written at a suitable reading level, they rarely cited references[41]. The concern here is that LLMs can produce plausible but incorrect information or 'hallucinations' which may not be obvious to non-expert readers, and may lead to misdiagnosis, excessive investigation, or adverse patient outcomes. If specialist LLMs are to be deployed for patient-facing use, methods for dealing with errors must be determined first.

Barriers to the deployment of LLMs

Regulation is struggling to keep up with the pace of LLM development – both the UK and EU AI acts have been through multiple iterations, demonstrating the challenges of regulating a dynamic field[42-44]. With general purpose LLMs, like ChatGPT, there is limited knowledge on the quality of the data that has been used in training, making their use in clinical practice particularly high risk[45]. Using LLMs in clinical decision-making also necessitates discussion regarding liability[46]. If a patient were to come to harm following LLM input in their care, there would need to be clear policy on who is held accountable for the decision. For example, if a chatbot consulted with undifferentiated patients, it is easy to see how LLMs could misinterpret a prompt with potentially life-threatening implications. Equally, patients must be aware of LLM use within their care, and it is uncertain whether this would actually increase patient concern. LLMs are also at risk of bias due to their dependence on their training data, which likely comes from high-income, Western countries[47]. It is therefore essential that models are trained and evaluated on diverse datasets prior to deployment.

A roadmap to LLM translation

Despite these issues, technologists continue to seek solutions to improve performance of LLMs. The risk of hallucination can be reduced through 'reinforcement learning with human feedback (RLHF),' where humans are directly involved in the training process[48], (for example, if the LLM outputs 'Humpty Dumpty sat on a fence', a human can correct the model to 'wall' and the LLM can adjust based on this feedback and improve on future iterations). Fine-tuning can also be used, which involves further training existing models on targeted datasets to enhance their capabilities carrying out specific tasks. LLMs are also intrinsically dependent on their training data. As clinical practice evolves with rapidly changing guidelines, technical solutions such as retrieval augmented generation (RAG) (a method which allows LLMs to search data outside its training set, for example clinical guidelines or patient datasets) could enhance trustworthiness of these models[49].

Technological limitations are likely to be resolved as newer iterations of models emerge, but regulatory issues must be solved if clinical translation of LLMs is to become a reality. Regulatory bodies must negotiate data quality assurance, model evaluation, and monitoring, and a clear definition of the scope of LLMs within medical practice[4,45]. This will likely require a multi-stakeholder approach including patients, healthcare professionals, software engineers, and policymakers.

Conclusion

LLMs are being adopted at a rapid pace and its use in healthcare is likely to expand in the coming years. Whilst its use may be reasonable in low-risk settings, it is critical that regulations are in place to ensure they are accurate and contributing meaningfully to the healthcare field. Barriers to deployment will likely be beyond technological barriers alone and include trustworthiness, confidentiality, and liability. This will likely require a multi-stakeholder approach including patients, healthcare professionals, software engineers and policymakers to establish codes of conduct, guidelines and standards for the use of LLMs within orthopaedic surgery. ■

References

References can be found online at www.boa.ac.uk/publications/JTO.